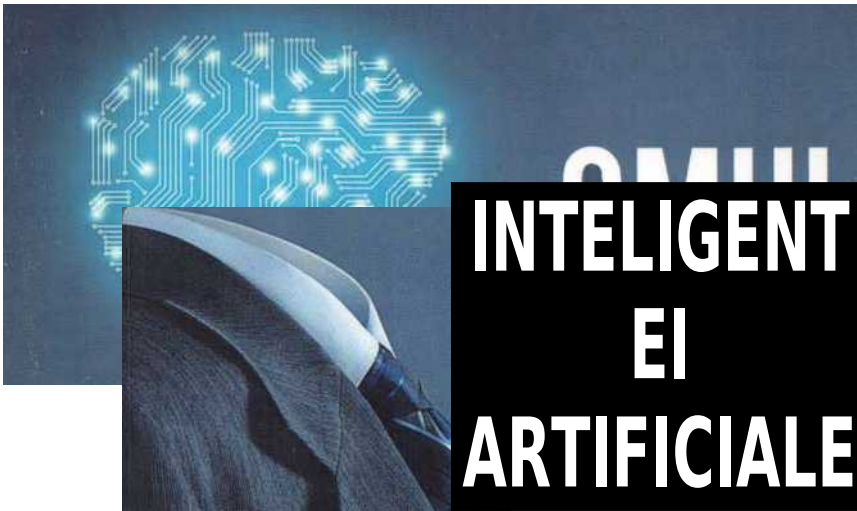
„Un ghid pasionant pentru provocările şi alegerile aduse de căutarea unui viitor măreţ al vieţii, inteligenţei şi conştiinţei – pe Pământ şi dincolo de el.“ i OK MUSK



INTELIGENT EI ARTIFICIALE

Una dintre cele mai spectaculoase evoluţii în cursul celor 13,8 miliarde de ani scurşi de la big bang este că acea materie mută şi neînsufleţită a devenit inteligentă. Cum s-a petrecut asta şi cum poate progresa inteligenţa în viitor? Ce are de spus ştiinţa despre istoria şi destinul inteligenţei în cosmosul nostru? —MAX TEGMARK .

VIATA 3.0

Max Tegmark (n. 1967) este un fizician american de origine suedeză care ţi-a manifestat încă din adolescenţă talentul de informatician, scriind în cod maşină un procesor de text pe care l-a vândut corporaţiei suedeze Dataindustrier AB. După ce a obţinut licenţa în fizică la Institutul Tehnologic Regal din Stockholm, a studiat în continuare fizica la Universitatea Berkeley din California, unde şi-a susţinut doctoratul. în prezent, este profesor la Institutul Tehnologic din Massachusetts (MIT). Cercetările sale din domeniul cosmologiei l-au condus către „ipoteza universului matematic" (prezentată în cartea sa *Our Mathematical* Univene), conform căreia toate structurile care există din punct de vedere matematic există şi fizic. Preocupat de inteligenţa artificială şi îngrijorat de perspectivele dezvoltării ei, Max Tegmark a înfiinţat Institutul pentru Viitorul Vieţii, prin care a reuşit să aducă în centrul atenţiei specialiştilor şi publicului larg probleme delicate legate de siguranţa inteligenţei artificiale.

MAX TEGMARK

VIAŢA 3.0

V

OMUL ÎN EPOCA INTELIGENŢEI ARTIFICIALE

Traducere din engleză
de Vlad Zografi

HUMANITAS
BUCUREŞTI

Coperta: Ioana Nedelcu
Tehnoredactor: Manuéla Măxineanu
Corectori: Cristina (elescu, Alexandru Anghel
DTP: Florina Vasiliu, Dan Dulgheru

Tipărit la Paper Prinţ – Brăila

Max Tegmark
Life 3.0: Being Humán in *file Age of Artificial Intelligence*
Copyright © 2017 by Max Tegmark
All rights reserved.

© HUMANITÁS, 2019, pentru prezenta versiune românească

Descrierea CIP a Bibliotecii Naţionale a României
Tegmark, Max
Viaţa 3.0: Omul în epoca inteligenţei artificiale *I* Max Tegmark; trad. din engleză de Vlad Zografi. -Bucureşti: Humanitás, 2019 ISBN 978-973-50-6402-0
I. Zografi, Vlad (trad.)
0048

EDITURA HUMANITÁS
Piaţa Presei Libere 1,023701 Bucureşti, România
tel. 021/408 83 50, fax 021/408 83 51
www.humanitas.ro

Comenzi online: www.libhumanitas.ro
Comenzi prin e-mail: vanzari@libhumanitas.ro
Comenzi telefonice: 021/3112330

Pentru Echipa FLI,

care ajacut totul
posibil

Mulţumiri

Le sunt sincer îndatorat tuturor celor care m-au încurajat şi m au ajutat să scriu această carte, între care:

familia mea, prietenii, profesorii, colegii şi colaboratorii, pentru sprijinul şi inspiraţia pe care mi le-au oferit de-a lungul anilor,

Mama, pentru că mi-a stârnit curiozitatea în privinţa conştiinţei şi a sensului,

Tata, pentru spiritul de luptător pus în slujba unei lumi mai bune, fiii mei Philip şi Alexander, pentru că m-au făcut să văd cum se naşte minunea inteligenţei umane,

toţi pasionaţîi de ştiinţă şi tehnologie de pretutindeni, care, de-a lungul anilor, prin întrebări şi comentarii m-au încurajat să-mi urmez şi să-mi public ideile,

agentul meu, John Brockman, care a insistat până am acceptat să scriu această carte,

Bob Penna, Jesse Thaler şi Jeremy England, pentru discuţiile lămuritoare despre quasari, sfaleroni şi, respectiv, termodinamică cei de la care an primit reacţii privind părţi din manuscris, între care Mama, fratele

meu Peter, Luisa Bahet, Rob Bensinger, Katerina Bergstrdm, Erik Brynjolfsson, Daniela Chita, David Chalmers, Nima Deghani, Henry Lin, Elin Malmskbld, Toby

Ord, Jeremy Owen, Lucas Perry, Anthony Romero, Nate Soares şi Jaan Tallinn,

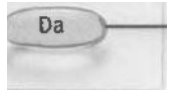
supereroii care au comentat versiuni integrale ale întregii cărţi, Meia, Tata, Anthony Aguine, Paul Almond, Matthew Graves, Phillip Helbig, Richard Mallah, David Marble, Howard Messig, Luino Seoane, Marin Solacic, editorul meu Dan Frank, iar mai presus de toate

Meia, iubita mea muză şi tovarăş de călătorie, pentru încurajarea, sprijinul şi inspiraţia pe care mi le-a oferit neîncetat şi fără de care această carte n-ar fi putut exista.

VIATA 3.0

V

Crezi că IA supraumană ar putea fi creată în acest secol?



Continuă cu pagina următoar

Sari la capitolul i
<P- 31)

PRELUDIU

Povestea Echipei Omega

Echipa Omega era sufletul companiei. Dacă restul întreprinderii aducea bani pentru a face ca lucrurile să funcţioneze, prin diverse aplicaţii comerciale ale inteligenţei artificiale în sens restrâns, Echipa Omega avansa în căutarea a ceea ce directorul general al companiei visase dintotdeauna: crearea inteligenţei artificiale generale. Majoritatea celorlalţi angajaţi îi priveau pe „omegi”, aşa cum îi numeau ei afectuos, ca pe nişte visători aflaţi mereu la decenii distanţă de scopul lor. îi simpatizau totuşi, fiindcă se bucurau de prestigiul pe care munca de pionierat a omegilor îl aducea companiei şi foloseau algoritmii îmbunătăţiţi pe care omegii li-i puneau la dispoziţie din când în când.

Ceea ce nu ştiau ei era ca omegii îşi făuriseră această imagine ca să ascundă un secret: se aflau extrem de aproape de înfăptuirea celui mai îndrăzneţ plan din istoria omenirii. Carismaticul director general îi alesese nu numai pentru că erau cercetători străluciţi, ci şi pentru ambiţia, idealismul şi dorinţa lor de a sluji omenirea. Le reamintea că planul lor era foarte primejdios, iar dacă guverne puternice l-ar afla, ar fi în

stare de orice – inclusiv să-i răpească – pentru a-i opri sau, mai curând, pentru a le fura codul. Ei erau însă dedicaţi cu trup şi suflet, cam din acelaşi motiv pentru care mulţi mari fizicieni ai lumii se alăturaseră Proiectului Manhattan de

construire a bombei atomice: erau convinşi că dacă n-o fac ei primii, o va face cineva mai puţin idealist.

Inteligenţa artificială creată de ei, numită Prometeu, se dezvolta continuu. Deşi capacităţile ei cognitive erau încă departe de cele ale oamenilor în multe domenii – în privinţa abilităţilor sociale, de pildă – omegii se străduiseră s-o perfecţioneze pentru o sarcină precisă: programarea sistemelor de inteligenţă artificială (IA). Aleseseră anume această strategie, urmând ideea matematicianului britanic Irving Good enunţată încă din 1965: „Definim o maşină ultrainteligentă ca pe o maşină care poate întrece în toate activităţile intelectuale orice om, oricât de deştept ar fi. Din moment ce proiectarea maşinilor este una dintre aceste activităţi intelectuale, o maşină ultrainteligentă ar putea proiecta maşini încă şi mai bune; ar avea loc atunci fără îndoială o «explozie a inteligenţei», iar inteligenţa umană ar fi lăsată mult în urmă. Aşadar, prima maşină ultrainteligentă ar fi ultima invenţie pe care omul ar avea nevoie s-o facă, cu condiţia ca maşina să fie suficient de docilă pentru a ne spune cum s-o ţinem sub control.“

Eî s-au gândit că, dacă ar putea iniţia această autoperfecţîonare recursivă, curând maşina ar deveni suficient de deşteaptă încât să deprindă singură toate celelalte abilităţi umane care i-ar fi folositoare.

PRIMELE MILIOANE

Era ora nouă dimineaţa, într-o zi de vineri, când s-au hotărât să-i dea drumul. Prometeu zumzăia în componentele lui electronice aşezate pe şiruri lungi într-o încăpere cu acces limitat şi aer condiţionat Din motive de siguranţă era complet decuplat de intemet, dar conţinea o copie locală a unei mari părţi a reţelei (Wildpedia, Biblioteca Congresului, Twitter, o selecţie

de pe Youtube, mare parte din Facebook etc.) pentru a-i folosi datele în procesul de învăţare.* Au ales acest moment de lansare ca să lucreze netulburaţi: rudele şi prietenii lor credeau că se află într-o casă de odihnă a companiei. Bucătăria era plină cu mâncare semipreparată şi băuturi energizante, iar ei erau pregătiţi să înceapă.

Când l-au lansat, Prometeu se descurca ceva mai slab decât ei la programarea sistemelor IA, dar compensa prin viteza uimitoare, făcând munca a o mie de oameni vreme de un an cât timp ei îşi beau paharul de Red Bull. Pe la to a.m. îşi încheiase prima repro- iectare a lui însuşi, vz.o, o versiune ceva mai bună, dar încă subumană. Când s-a lansat Prometeu 5.0 la 2 p.m., omegii au fost însă uluiţi: le depăşise cu mult performanţa, iar ritmul progresului părea să se accelereze. La căderea serii s-au hotărât să-l angajeze pe Prometeu 10.0 în faza a doua a planului lor: să câştige bani.

Prima lor ţintă a fost MTurk, Amazon Mechanical Turk. După lansarea lui în 2005 ca piaţă de externalizare prin intemet, a crescut rapid, zeci de mii de oameni din lumea întreagă concurând anonim douăzeci şi patru de ore pe zi pentru a îndeplini activităţi structurate numite HITS, „Human Intelligence Tasks“ (sarcini care presupun inteligenţă umană). Spectrul acestor sarcini se întindea de la transcrierea înregistrărilor audio până la clasificarea imaginilor şi redactarea descrierilor de pagini web, iar toate aveau ceva în comun: dacă reuşeai să le îndeplineşti, nimeni nu şi-ar fi dat seama că eşti o IA. Prometeu ro.o a

* Pentru simplitate, am presupus în povestea mea că avem de-a face cu aceeaşi economie ţi tehnologie de azi, deşi majoritatea cercetătorilor cred că de IA generală la nivel uman ne despart cel puţin decenii. Planul Omega ar fi şi mai uşor de realizat în viitor dacă economia digitală şi-ar continua creşterea şi ar fi disponibile online şi mai multe servicii, fără a se cere informaţii în prealabil. (N. o.)

îndeplinit în mod acceptabil cam jumătate din categoriile de sarcini. Pentru fiecare asemenea categorie de sarcini, omegii l-au pus pe Prometeu să proiecteze un modul IA limitat, destinat numai acestor sarcini. Au încărcat apoi modulul pe Amazon Web Services, o platformă de *cloud-computing*† pe care puteau rula atâtea maşini virtuale câte erau închiriate. La fiecare dolar plătit diviziei de *cloud-computing* de Ia Amazon, câştigau mai mult de doi dolari de la divizia MTurk a aceluiaşi Amazon. Amazon habar n avea că exista asemenea posibilitate de speculă chiar în interiorul propriei companii!

Pentru a-şi ascunde urmele, omegii creaseră discret în lunile precedente mii de conturi MTurk în numele unor persoane fictive, iar modulele Prometeu îşi asumau acum identităţile acestora. Clienţii MTurk erau plătiţi de regulă după vreo opt ore, moment în care omegii reinvesteau banii în mai mult timp de *cloud-computing,* folosind module ameliorate create de ultima versiune a lui Prometeu, aflat în continuă perfecţionare. Pentru că îşi puteau dubla banii la fiecare opt ore, curând au saturat fondurile MTurk, şi au constatat că nu pot câştiga mai mult de un milion de dolari pe zi fără să atragă atenţia asupra lor. Dar asta era mai mult decât suficient pentru a-şi finanţa următorul pas, eliminând orice nevoie de a cere bani directorului financiar-contabil.

JOCURI PERICULOASE

† *Cloud-computing* („calculare în nor") presupune utilizarea puterii de calcul sau de stocare a serverelor aflate la distanţă prin intermediul intemetului, acele servere fiind închiriate la cerere. *(N. t)*

Dincolo de descoperirile lor în IA, omegii se amuzau să-şi imagineze cum pot câştiga bani cât mai rapid, odată lansat Prometeu. în fond, întreaga economie digitală le stătea la dispoziţie, dar nu cumva era mai bine să înceapă să creeze jocuri pe calculator, muzică, filme sau software, să scrie cărţi sau articole, să tranzacţio- neze acţiuni la bursă sau să facă invenţii şi să le vândă? Asta ar însemna să maximizeze rata de recuperare a investiţiei, dar strategiile normale de investiţie erau o parodie derulată cu încetinitorul a ceea ce puteau face ei: în vreme ce un investitor normal ar fi mulţumit cu un beneficiu de 9% pe *an,* investiţiile lor MTurk produceau 9% pe *oră,* generând de opt ori mai mulţi bani în fiecare zi. Aşa încât, după saturarea MTurk, ce urma?

S-au gândit mai întâi să distrugă bursa – la urma urmei, mulţi dintre ei refuzaseră un job bine plătit în dezvoltarea IA pentru fonduri speculative, care investeau sume mari exact în aceeaşi idee. Unii îşi aminteau că în felul ăsta câştigase IA primele milioane în filmul *Transcendence.* Dar noile reglementări privind

instrumentele financiare derivate, după criza din anul precedent, le limitau alegerile. Au înţeles curând că, deşi puteau avea beneficii mult mai mari decât alţi investitori, era puţin probabil să obţină câştiguri comparabile cu ce le-ar aduce vânzarea propriilor produse. Când ai la dispoziţie prima maşinărie superinteligentă din lume, e de preferat să investeşti în propriile tale companii decât în companiile altora! Deşi puteau apărea unele excepţii (cum ar fi folosirea talentelor supraumane de hacker ale lui Prometeu pentru a căpăta informaţii din interior şi a cumpăra acţiuni care urmau să crească), omegii şi-au spus că nu merita să atragă atenţia asupra lor.

Când s-au concentrat asupra produselor pe care le puteau dezvolta şi vinde, jocurile pe calculator li s-au părut cea mai bună alegere. Prometeu se putea specializa rapid în proiectarea unor jocuri atrăgătoare, manevrând cu uşurinţă codificarea, grafica, formarea imaginilor şi toate celelalte sarcini necesare pentru a obţine un produs finit. Mai mult, după digerarea tuturor datelor din reţea despre preferinţele oamenilor, ar şti exact ce place fiecărei categorii de gameri, şi ar dobândi o abilitate supraumană de a optimiza un joc pentru câştigul maxim. *The Elder Scrolls* V; *Skyrirn,* un joc cu care omegii au pierdut mai mult timp decât erau dispuşi sa recunoască, adusese câştiguri de peste 400 de milioane de dolari în prima săptămână de la lansare, în 201 r, iar ei erau convinşi că Prometeu putea crea ceva cel puţin Ia fel de tentant în douăzeci şi patru de ore, folosind resurse de *ctoud-computmg* costând un milion de dolari. L-ar putea vinde apoi online şi l-ar pune pe Prometeu să creeze personaje umane care vorbesc despre joc în blogosferă. Dacă asta le-ar aduce 250 de milioane de dolari într-o săptămână, ei şi-ar putea dubla investiţia de opt ori în opt zile, cu un beneficiu de 3% pe oră – mai mic decât cel de la început, cu MTurk, dar

mult mai sigur. Dezvoltând o serie de alte jocuri în fiecare zi, şi-au făcut socoteala că ar putea curând câştiga 10 miliarde de dolari, fără a satura piaţa jocurilor.

Dar o specialistă în cybersecuritate din echipa lor a analizat acest plan al jocurilor. Ea a arătat că ar apărea un risc inacceptabil dacă Prometeu *ar evada* şi ar căpăta control asupra propriului

destin. Fiindcă nu puteau şti cum ar evolua obiectivele lui în decursul autoperfecţionării recursive, au hotărât să fie prevăzători şi să-l menţină „închis“ pe Prometeu, ca să nu evadeze pe intemet. Pentru maşinăria principală care rula în încăperea cu acces limitat, ei foloseau izolarea fizică: nu exista conexiune intemet, iar singurul output de la Prometeu era sub forma unor mesaje şi documente pe care le trimitea unui calculator controlat de omegi.

însă pe un calculator conectat la intemet, rularea oricărui program complicat creat de Prometeu era o propunere riscantă: cum omegii nu-şi dădeau seama până la capăt ce-ar face, nu puteau şti dacă nu cumva ar începe, de pildă, să se răspândească viral online. Când au testat software-ul scris de Prometeu pentru sarcinile MTurk, s-au protejat de acest risc rulându-l doar în interiorul unei maşini virtuale. Acesta e un program care simulează un calculator: de exemplu, mulţi utilizatori de Mac cumpără un software maşină virtuală care le permite să ruleze programele Windows făcându-le să creadă că e o maşină Windows veritabilă. Omegii îşi creaseră propria lor maşină virtuală, botezată Cutia Pandorei, care simula o maşină ultrasimplificată, lipsită de toate accesoriile pe care le asociem de regulă cu calculatoarele: nici un fel de tastatură, monitor, difuzoare, legătură intemet – nimic. Pentru transcrierile audio MTurk, omegii au aranjat aşa încât tot ce putea intra în Cutia Pandorei să fie un singur fişier audio, şi tot ce putea ieşi să fie un singur document text – transcrierea. Aceste „legi ale cutiei“ erau pentru software-ul dinăuntru cam ce sunt legile fizicii pentru noi, în interiorul universului nostru: softwa- re-ul nu putea ieşi din cutie, la fel cum noi nu putem călători cu viteze mai mari decât viteza luminii, oricât de deştepţi am fi. Cu excepţia acelor unice input şi output, software-ul dinăuntrul Cutiei Pandorei era efectiv prizonier într-un univers paralel, cu propriile lui reguli

computaţionale. Omegii se temeau atât de mult de evadare, încât au adăugat şi o restricţie temporală, limitând timpul de viaţă al codurilor în care n-aveau încredere. De pildă, de fiecare dată când software-ul de transcriere termina de transcris un fişier audio, întreaga memorie a Cutiei Pandorei era automat ştearsă, iar programul era reinstalat de la zero. Astfel, când începea o nouă sarcină de transcriere, nu ştia ce se întâmplase înainte, deci nu putea învăţa în cursul timpului.

Când omegii au folosit norul Amazon pentru proiectul lor MTurk, ei au putut pune toate modulele create de Prometeu pentru îndeplinirea sarcinilor în asemenea cutii virtuale din nor, pentru că input-urile şi output-urile MTurk erau foarte simple. Metoda nu mai funcţiona însă la jocurile pe calculator cu multă grafică, care nu puteau fi „închise în cutie“ fiindcă trebuiau să aibă acces deplin la întregul hardware al calculatorului jucătorului. Mai mult, ei n-au vrut să se expună riscului ca vreun utilizator expert să le analizeze codul jocului, să descopere Cutia Pandorei şi să vrea să vadă ce-i înăuntrul ei. Riscul evadării scotea din calcul nu doar piaţa jocurilor, dar şi piaţa extrem de profitabilă a altor tipuri de software, cu sute de miliarde de dolari care puteau fi câştigaţi.

PRIMELE MILIARDE

Omegii se concentraseră asupra produselor foarte scumpe, pur digitale (evitând o fabricare lentă) şi uşor de înţeles (de pildă, texte sau filme despre care ştiau că nu prezintă riscul evadării). în cele din urmă se hotărâseră să lanseze o companie media şi să înceapă cu desene animate. Website-ul, planul de marketing şi comunicatele de presă fuseseră deja gata înainte ca Prometeu să devină superinteligent – nu lipsea decât

conţinutul

Deşi duminică dimineaţa capacităţile lui Prometeu deveniseră uimitoare şi se extrăgeau constant bani de la MTurk, abilităţile lui intelectuale rămăseseră limitate: Prometeu fusese în mod deliberat optimizat pentru a proiecta sisteme IA şi a scrie programe care executau sarcini plictisitoare pentru MTurk. De exemplu, nu se pricepea să facă filme, iar asta nu din vreun motiv profund, ci din acelaşi motiv pentru care James Cameron nu se pricepea să facă filme atunci când s-a născut: este o abilitate pe care o deprinzi în timp. La felea un copil uman, Prometeu putea învăţa orice voia din datele la care avea acces. Dacă lui Cameron i-a luat ani de zile ca să înveţe să scrie şi citească, Prometeu a reuşit asta de vineri, când a avut timp şi să citească întreaga Wikipedie şi câteva milioane de cărţi. Să faci filme era însă mai greu. A scrie un scenariu pe care oamenii să-l găsească interesant era la fel de greu ca a scrie o carte, cerând înţelegerea detaliată a societăţii umane şi a ceea ce îi atrage pe oameni. Transformarea scenariului într-un fişier video presupunea o cantitate enormă de calcule pentru formarea imaginilor cu actorii simulaţi şi cu decorul complex prin care se mişcau, voci simulate, o coloană sonoră potrivită etc. Duminică dimineaţa Prometeu a urmărit un film de două ore într-un minut, timp în care a citit şi cărţile pe care se baza filmul şi toate cronicile şi evaluările online. Omegii au observat că, după ce Prometeu a vizionat câteva sute de filme, a început să prevadă corect ce gen de cronici s-ar putea scrie despre un film şi cum ar fi apreciat de diferite tipuri de public. A învăţat să scrie cronici pătrunzătoare, comentând totul, de la intrigă şi jocul actorilor Ia detalii tehnice cum ar fi lumina şi unghiul de filmare. Omegii au dedus de aici că, atunci când îşi va face propriile filme, Prometeu va şti ce înseamnă succesul.

Omegii l-au învăţat pe Prometeu să înceapă cu animaţia, fără să se întrebe cine sunt actorii simulaţi. Duminică noaptea şi-au încheiat furtunosul weekend înarmându-se cu bere şi popcom, stingând luminile şi urmărind filmul de debut al lui Prometeu. Era o comedie-fantezie animată, în genul lui *Frozen,* iar formarea imaginilor fusese realizată cu un cod „închis“ al lui Prometeu rulat în norul Amazon, cheltuindu-se astfel aproape tot milionul de dolari câştigat în acea zi de la MTurk. Când a început filmul, ome- gilor li s-a părut deopotrivă fascinant şi înspăimântător faptul că era creaţia unei maşini care nu fusese îndrumată de om. Curând însă, râdeau la poante şi îşi ţineau respiraţia în momentele drama tice. Unii chiar au lăcrimat la sfârşitul emoţionant, atât de cufundaţi în această realitate ficţională, încât au uitat cine e autorul.

Omegii au plănuit ca lansarea site-ului să aibă loc vineri, lăsând timp lui Prometeu să producă mai mult conţinut şi lor să se ocupe de alte lucruri: să cumpere publicitate şi să recruteze angajaţi pentru companiile fictive pe care le înfiinţaseră în ultimele luni. Pentru a-şi ascunde urmele, povestea oficială era că propria lor companie media (fără vreo legătură directă cu ei) a cumpărat cea mai mare parte a conţinutului de la producători de film indepen denţi, de regulă startup-uri de înaltă tehnologie din zone sărace. Aceşti falşi furnizori au fost plasaţi în locuri îndepărtate, cum ar fi Tiruchchirappalli şi lakuţk, pe care nici cei mai curioşi jurnalişti nu s-ar osteni să le viziteze. Singurii angajaţi pe care i-au recrutat efectiv lucrau în marketing şi administraţie, şi oricui ar fi pus întrebări le-ar fi răspuns că echipa lor de producţie se afla în altă parte, iar deocamdată nu dădea interviuri. Pentru a se potrivi cu povestea oficială, au ales drept slogan al companiei „Să canalizăm talentul creator al lumii“, şi au ales ca brand ideea că sunt radical diferiţi prin folosirea tehnologiei de vârf pentru a-i ajuta pe oamenii creativi, în special pe

cei din ţările în curs de dezvoltare.

Vineri, vizitatorii curioşi care au intrat pe site-ul lor au dat peste ceva ce amintea de serviciile Netflix şi Hulu, dar cu o diferenţă semnificativă: toate serialele animate erau noi Păreau captivante – majoritatea serialelor constau în episoade de 45 de minute, cu un scenariu bine articulat, fiecare încheindu-se într-un fel care te făcea nerăbdător să afli ce se întâmplă în episodul următor. Şi erau mai ieftine decât cele produse de concurenţă. Primul episod al fiecărui serial era gratuit, şi le puteai vedea pe celelalte plătind 49 de cenţi pe episod, cu reduceri pentru întregul serial. La început existau doar trei seriale cu trei episoade fiecare, dar zilnic erau adăugate noi episoade, satisfăcând diverse gusturi. în primele două săptămâni ale lui Prometeu, talentele lui de cineast au sporit rapid, nu numai în privinţa calităţii filmelor, dar şi a algoritmilor pentru simularea personajelor şi formarea imaginilor, ceea ce a redus mult costul pentru *ctoud-computing* la fiecare nou episod. Ca atare, omegii au putut lansa zeci de noi seriale în prima lună, publi- cul-ţintă cuprinzând toate categoriile, de la copii la adulţi, şi s-au extins pe toate pieţele importante din lume corespunzând limbilor de largă circulaţie, ceea ce a dat site-ului un caracter internaţional incomparabil cu al concurenţilor. Unii comentatori au fost impresionaţi de faptul că nu doar banda sonoră era în mai multe limbi, dar şi imaginile înseşi: de pildă, când un personaj vorbea în italiană, mişcarea buzelor se potrivea cuvintelor italieneşti, la fel şi gestica tipic italiană. Deşi Prometeu putea crea acum filme cu actori simulaţi imposibil de deosebit de oamenii reali, omegii n-au făcut-o, pentru ca nu cumva să se dea de gol. Au lansat însă multe seriale cu personaje umane animate semirealiste, în genul show-urilor TV şi filmelor tradiţionale.

Reţeaua lor s-a bucurat de mare succes, iar numărul vizionărilor a crescut spectaculos. Mulţi fani au găsit că

personajele şi scenariile sunt mai inteligente şi mai interesante decât în cele mai scumpe producţii hollywoodiene pentru marele ecran, şi erau încântaţi ca le pot urmări la preţuri rezonabile. Beneficiind de o publicitate agresivă (omegii şi-o permiteau fiindcă producţia nu-i costa aproape nimic), de o acoperire media excelentă şi cronici entuziaste, după o lună de la lansare venitul lor global a ajuns la zece milioane de dolari pe zi. Peste două luni depăşiseră Netflix, iar peste trei au ajuns la o sută de milioane de dolari pe zi, începând să rivalizeze cu Time Wamer, Disney, Comcast şi Fox, ca unul dintre cele mai mari imperii media din lume.

Acest senzaţional succes a atras atenţia asupra lor, s-a spus că ar poseda o IA puternică, însă folosind doar o mică parte din venit, omegii au desfăşurat o reuşită campanie de dezinformare. Din- tr-un sediu nou şi kitschos din Manhattan, purtători de cuvânt proaspăt angajaţi relatau în amănunt poveştile lorinventate Mulţi oameni au fost angajaţi în chip de diversiune, între care scenarişti adevăraţi din lumea întreagă, pentru a elabora noi seriale, fără ca vreunul să aibă habar de Prometeu. Derutanta reţea internaţională de subcontractori a făcut cu putinţă ca majoritatea angajaţilor să presupună că alţii de altundeva fac cea mai mare parte din muncă.

Pentru a deveni mai puţin vulnerabili şi a evita bănuielile provocate de excesul de *cloud computing,* au angajat de asemenea ingineri ca să construiască o serie de centre puternice de calculatoare pretutindeni în lume, deţinute de companii fictive aparent neafiliate. Deşi numite „centre verzi de date“, fiindcă erau alimentate în mare parte cu energie solară, acestea se concentrau de fapt asupra calculului, nu asupra stocării. Prometeu le proiectase până în cele mai mici detalii, folosind doar hardware-ul disponibil pe piaţă şi optimizându-le pentru a minimiza timpul de construcţie.

Cei care le-au construit şi Ie conduceau habar n-aveau ce se calcula acolo, îşi închipuiau că au de-a face cu instalaţii comerciale de *cloud-computing,* asemănătoare celor deţinute de Amazon, Google şi Microsoft, şi ştiau doar că toate vânzările sunt gestionate de undeva de departe.

NOI TEHNOLOGII

Pe măsură ce treceau lunile, imperiul de afaceri condus de omegi pătrundea în tot mai multe domenii ale economiei mondiale, gra ţie planificării supraumane a lui Prometeu. Analizând atent datele internaţionale, încă din prima săptămână le oferise omegilor un plan detaliat de creştere pas cu pas, plan continuu îmbunătăţit odată cu acumularea datelor şi resurselor de calcul. Deşi Prometeu era departe de a fi atotcunoscător, capacităţile lui depăşiseră atât de mult limitele omenescului, încât omegii îl priveau ca pe oracolul perfect, oferind supus răspunsuri şi sfaturi strălucite la toate întrebările lor.

Software-ul lui Prometeu era acum optimizat pentru a obţine maximumul din bardul mai curând mediocru, inventat de oameni, pe care rula, şi, după cum anticipaseră omegii, Prometeu a găsit căi de a îmbunătăţi spectaculos acest hardware. Temându se de o evadare, ei au refuzat construcţia robotică a unor dispozitive pe care Prometeu ar fi controlat-o direct. Au angajat în schimb numeroşi oameni de ştiinţă şi ingineri de înaltă clasă din diferite colţuri ale lumii şi le-au oferit rapoartele interne de cercetare scrise de Prometeu, pretinzând că erau ale unor cercetători din alte locuri. Aceste rapoarte prezentau în detaliu noi efecte fizice şi tehnici de producţie, pe care curând inginerii lor le-au testat, înţeles şi stăpânit. Desigur, ciclurile umane normale de cercetare şi dezvoltare durează ani, în mare

măsură fiindcă ele presupun multe cicluri lente de încercări şi eşecuri. Acum situaţia era diferită: Prometeu cunoştea deja următorii paşi de urmat, aşa încât erau limitaţi doar de viteza cu care oamenii puteau fi îndrumaţi să înţeleagă şi să construiască ce trebuia. Un profesor bun îi poate ajuta pe studenţi să înveţe ştiinţa mult mai repede decât dacă ei ar descoperi totul pe cont propriu pornind de la zero, iar Prometeu a făcut pe ascuns acelaşi lucru cu aceşti cercetători. Cum Prometeu putea prezice cu exactitate cât le-ar lua oamenilor să înţeleagă şi să construiască dispozitive cu diferite instrumente, a găsit cea mai scurtă cale de înaintare, dând prioritate noilor instrumente care ar fi imediat înţelese şi construite, şi care ar ajuta la dezvoltarea unor instrumente şi mai avansate.

Echipele de ingineri au fost astfel încurajate să-şi folosească propriile maşini pentru a construi unele mai bune. Această autoapro- vizionare nu numai că a adus economii băneşti, dar i-a făcut pe omegi mai puţin vulnerabili la viitoare ameninţări din lumea exterioară. în doi ani, produceau un hardware incomparabil mai bun decât existase vreodată în lume. Pentru a nu-şi ajuta concurenţa, au păstrat secretă această tehnologie, folosind-o doar pentru a-l upgrada pe Prometeu.

Ce a observat totuşi lumea a fost o uimitoare explozie tehnologică. Companii nou înfiinţate de pretutindeni lansau noi produse revoluţionare aproape în toate domeniile. Un startup sud-coreean a lansat un nou acumulator care stoca de două ori mai multă energie decât acumulatorul de la laptopul tău, fiind de două ori mai uşor, şi se încărca în mai puţin de un minut. O firmă finlandeză a venit cu un panou solar ieftin de două ori mai eficient decât al concurenţilor. O companie germană a anunţat un nou tip de cablu care putea fi produs pe scară largă şi era supraconductor la temperatura camerei, revoluţionând sectorul energetic. Un grup de cercetare în biotehnologie cu sediul la Boston a anunţat

că se afla în faza a doua a testelor clinice primul medicament eficace pentru slăbit, fără efecte secundare, şi se zvonea că un producător indian vindea deja aşa ceva pe piaţa neagră. 0 companie califomiană a ripostat anunţând faza a doua a testelor pentru un medicament-minune anticancer, care ajută sistemul imunitar să recunoască şi să atace celulele care prezintă unele dintre cele mai frecvente mutaţii cancerigene. Exemplele veneau în cascadă, întărind ideea că era vorba de o nouă epocă de aur a ştiinţei. Pe de altă parte, companii de robotică răsăreau pretutindeni în lume ca ciupercile. Nici unul dintre roboţi nu se apropia de inteligenţa umană, iar cei mai mulţi dintre ei nu semănau deloc cu oamenii. Ei au perturbat însă violent economia, înlocuindu-i treptat pe cei mai mulţi lucrători din industrie, transport, comerţul cu amănuntul şi cu ridicata, construcţii, minerit, agricultură, silvicultură şi pescuit.

Ce nu s-a observat, graţie eforturilor unei echipe de avocaţi experţi, a fost că toate aceste firme erau controlate prin intermediari de omegi. Prometeu invada birourile de brevete cu invenţii senzaţionale prin diferiţi reprezentanţi, iar aceste invenţii au condus treptat la dominarea tuturor sectoarelor tehnologiei.

Deşi noile companii şi-au făcut inamici puternici în rândul concurenţei, ele şi-au făcut încă şi mai mulţi prieteni puternici. Erau extraordinar de profitabile, şi sub sloganuri de genul „Investim în comunitatea noastră“ cheltuiau părţi însemnate din profit angajând oameni pentru proiecte comunitare – adesea aceiaşi oameni care fuseseră concediaţi de companiile rivale aduse în pragul falimentului. Omegii foloseau analize ale lui Prometeu ca să identifice joburi optime pentru angajaţi şi comunitate, cu costuri minime, adaptate condiţiilor locale. în regiunile cu un nivel de trai ridicat se concentrau adesea asupra întării comunităţii, familiei

şi culturii, iar în regiunile sărace se preocupau în plus de înfiinţarea şi întreţinerea şcolilor, sănătate, ajutorarea persoanelor în vârstă, locuinţe la preţuri modice, spaţii verzi şi infrastructura de bază. Mai peste tot cetăţenii erau de părere că asemenea lucruri trebuiau făcute de multă vreme. Politicienii locali primeau donaţii generoase şi încurajau aceste investiţii comunitare.

DOBÂNDIREA PUTERII

Omegii creaseră o companie media nu doar pentru a-şi finanţa primele întreprinderi tehnologice, dar şi în vederea următorului pas al ambiţiosului lor plan: stăpânirea lumii. La un an după lansare, au înfiinţat pretutindeni în lume canale de ştiri remarcabil de bune. Spre deosebire de celelalte canale, ele au fost în mod deliberat concepute pentru a merge în pierdere şi au fost considerate un serviciu public. De fapt, canalele lor de ştiri nu produceau nici un fel de venit: nu dădeau publicitate şi erau accesibile gratuit oricui avea conexiune intemet. Restul imperiului lor media aducea atâtea câştiguri, încât îşi permiteau să cheltuiască pe serviciul de ştiri mult mai mult decât orice întreprindere jurnalistică din istorie – iar roadele s-au văzut. Prin recrutare agresiva, oferind salarii extrem de competitive ziariştilor şi reporterilor de investigaţie, au făcut dezvăluiri senzaţionale cu un remarcabil talent. Folosind serviciile unei reţele globale care plătea pe oricine venea cu o ştire interesantă, de la corupţia locală până la un eveniment emoţionant, ei erau de regulă primii care o anunţau Cel puţin asta credea lumea: de fapt, adesea ei erau primii fiindcă relatările atribuite jurnaliştilor fuseseră descoperite de Prometeu prin monitorizarea în timp real a intemetului. Pe toate aceste

site-uri video de ştiri apăreau de asemenea articole.

Prima fază a strategiei legate de ştiri presupunea câştigarea încrederii oamenilor, ceea ce au reuşit cu mare succes. Acceptând să piardă bani, au obţinut o remarcabilă acoperire locală şi regională, iar ziariştii de investigaţie dezvăluiau deseori scandaluri care ţineau trează atenţia publicului. Dacă într-o ţară divizarea politică era accentuată şi lumea se obişnuise cu ştiri partizane, ei lansau câte un canal de ştiri pentru fiecare facţiune, deţinut chipurile mereu de câte o alta companie, câştigând treptat încrederea acelei facţiuni. Unde era posibil, apelau la intermediari pentru a cumpăra cele mai influente canale existente, iar apoi le îmbunătăţeau eliminând publicitatea şi introducând propriul lor conţinut, în ţările în care cenzura şi ingerinţele politicului le ameninţau eforturile, acceptau la început orice le cerea guvernul pentru a rezista, cu sloganul intern secret „Adevărul şi numai adevărul, însă poate nu întreg adevărul“. De regulă, Prometeu oferea sfaturi excelente în asemenea situaţii, stabilind ce politicieni trebuie prezentaţi favorabil şi ce politicieni (în general corupţii locali) pot fi expuşi. Prometeu le spunea de asemenea ce sfori trebuie trase, cine şi cum trebuie mituit.

Această strategie a avut un succes zdrobitor pretutindeni în lume, canalele controlate de omegi devenind cele mai de încredere surse de ştiri. Chiar şi în ţările unde guvernele încercau să contracareze creşterea popularităţii lor şi-au dobândit credibilitate, iar multe din ştirile lor se răspândeau pe căi neoficiale. Directorii canalelor de ştiri concurente simţeau că bătălia era pierdută: cum poţi să obţii profit în competiţie cu cineva mai bine finanţat, care îşi oferă produsele gratis? Cu audienţa în scădere, tot mai multe reţele hotărau să-şi vândă canalele de ştiri – de regulă vreunui consorţiu care se dovedea apoi că e controlat de omegi.

La vreo doi ani de la lansarea lui Prometeu, când faza de câştigare a încrederii era în mare parte încheiată, omegri au iniţiat faza a doua a strategiei lor legate de ştiri: persuasiunea. încă dinainte observatorii fini au sesizat indicii ale agendei politice aflate în spatele noilor medii de informare: părea să existe o uşoară tendinţă spre centru, departe de extremismele de tot felul. Mulţimea lor de canale orientate către diferite grupuri reflecta în continuare animozitatea dintre Statele Unite şi Rusia, India şi Pakistan, diverse religii, facţiuni politice etc., dar tonul critic era uşor îndulcit, în prim-plan aflându-se temele concrete implicând banii şi puterea, iar nu atacurile la persoană, răspândirea panicii şi zvonurile fără acoperire. Odată începută cu adevărat faza a doua, această tendinţă de dezamorsare a vechilor conflicte a devenit tot mai vizibilă, prezentându-se frecvent relatări despre înţelegerile dintre adversarii tradiţionali, împreună cu investigaţii care arătau că mulţi susţinători vocali ai conflictelor erau mânaţi de interese materiale personale.

Comentatorii politici au observat că, în paralel cu atenuarea conflictelor regionale, părea să apară o presiune concertată către reducerea ameninţărilor globale. De pildă, riscurile unui război nuclear au început brusc să fie discutate pretutindeni. Câteva filme de mare succes se bazau pe scenarii în care un război nuclear era declanşat din întâmplare sau în mod deliberat, prezentând urmările lui distopice – iama nucleară, prăbuşirea infrastructurii, foametea generalizată. Documentare abil concepute arătau ce impact ar avea iama nucleară asupra fiecărei ţări. Oamenilor de ştiinţă şi politicienilor care susţineau dezangajarea nucleară li se acorda un timp amplu de emisie, în special pentru a discuta rezultatele unor noi studii privind măsurile ce pot fi luate – studii finanţate de organizaţii ştiinţifice care primiseră donaţii generoase de la noile companii de tehnologie. Ca atare,

a apărut tendinţa politică de a se depăşi starea de alertă şi a reduce arsenalele nucleare. Mediile de informare acordau de asemenea o atenţie sporită modificării globale a climei, subliniind adesea inovaţiile tehnologice datorate lui Prometeu, care scăzuseră mult preţul energiei regenerabile, şi încurajând guvernele să investească in asemenea noi infrastructuri energetice.

În paralel cu preluarea controlului asupra mediilor de informare, omegii l-au folosit pe Prometeu pentru a revoluţiona educaţia. Pornind de la cunoştinţele şi înzestrările fiecăruia în parte, Prometeu putea găsi calea cea mai rapidă de învăţare a unei materii noi, aşa încât fiecare să fie atras şi motivat să persevereze, şi putea produce materiale video, texte, exerciţii şi alte instrumente didactice optimizate. Companiile controlate de omegi au lansat deci pe piaţă cursuri online despre aproape orice, adaptate nu doar mediului lingvistic şi cultural, dar şi nivelului de pornire al elevului. Indiferent dacă erai un analfabet de 43 de ani care voia să înveţe să citească sau un doctor în biologie interesat de ultimele noutăţi în imunoterapia cancerului, Prometeu avea pentru tine cursul ideal. Aceste oferte nu semănau deloc cu celelalte cursuri online: profitând de talentele de cineast ale lui Prometeu, segmentele video te incitau cu adevărat prin metafore sugestive cu efect direct asupra ta, facându-te să-ţi doreşti să afli mai multe. Unele cursuri erau contra cost, dar multe erau gratuite, spre încântarea profesorilor din lumea întreagă care le puteau folosi la şcoală – şi a oricui era doritor să înveţe ceva.

Aceste supraputeri educaţionale s-au dovedit a fi instrumente eficiente în scopurile politice, creând „seriale de persuasiune“, în care dezvăluirile dintr-un film te făceau să-ţi revizuieşti opiniile şi te îndemnau să priveşti un alt film despre un subiect înrudit, pesemne la fel de convingător. Când scopul era, de pildă, dezamorsarea unui conflict între două naţiuni, în ambele ţări

erau difuzate independent documentare istorice care prezentau originea şi evoluţia conflictului într-o lumină mai nuanţată. Ştiri cu caracter pedagogic îi demascau pe aceia din propria tabără care trăgeau beneficii din perpetuarea conflictului şi prezentau tehnicile prin care ei îl alimentau. În acelaşi timp, figuri agreabile din cealaltă naţiune începeau să apară în emisiuni populare ale canalelor de divertisment, la fel cum în trecut personaje prezentate într-o lumină favorabilă au susţinut mişcările de emancipare civică şi cele pentru drepturile persoanelor gay.

Curând, comentatorii politici au observat creşterea sprijinului în favoarea unei agende concentrate în jurul a şapte sloganuri:

1. Democraţie
2. Reduceri fiscale
3. Reduceri ale serviciilor sociale guvernamentale
4. Reducerea cheltuielilor militare
5. Comerţ liber
6. Graniţe deschise
7. Companii responsabile social

Mai puţin evident era scopul subteran: erodarea tuturor structurilor de putere din lume. Punctele 2-7 erodau puterea statului, iar democratizarea lumii dădea imperiului de afaceri al omegilor o mai mare influenţă în selectarea liderilor politici. Companiile responsabile social slăbeau puterea statului preluând tot mai multe servicii pe care guvernele Ie ofereau (sau ar fi trebuit să le ofere). Elita de afaceri tradiţională era slăbită pur şi simplu pentru că nu putea concura pe piaţa liberă cu companiile susţinute de Prometeu, aşa încât deţinea o parte tot mai mică din economia mondială. Liderilor de opinie tradiţionali, de la partide politice la grupări religioase, le lipsea mecanismul de persuasiune pentru a face faţă imperiului media al omegilor.

Aşa cum se întâmplă în orice schimbare radicală,

existau învingători şi învinşi Deşi în majoritatea ţărilor se simţea o undă de optimism, odată ce educaţia, serviciile sociale şi infrastructura se îmbunătăţiseră, conflictele se domoliseră, iar companiile lansau tehnologii inovatoare care schimbau vieţile oamenilor, nu toată lumea era fericită. în vreme ce mulţi muncitori concediaţi erau reangajaţi pentru proiecte comunitare, cei care deţinuseră putere şi bogăţie au văzut în general că ambele se topesc. Fenomenul a început în sectoarele media şi tehnologic, dar s-a răspândit aproape pretutindeni. Diminuarea conflictelor a provocat reduceri ale bugetelor de apărare care i-au atins pe furnizorii militari. De regulă, companiile nou înfiinţate nu erau cotate la bursă, cu justificarea că acţionarii care vor să obţină profit maxim ar obstrucţiona proiectele comunitare. Prin urmare, bursa s-a devalorizat la nivel mondial, ameninţându-i deopotrivă pe magnaţii finanţelor şi pe cetăţenii de rând care contau pe fondurile lor de pensii. De parcă scăderea profiturilor companiilor listate la bursă n-ar fi fost de ajuns, firmele de investiţii din lumea întreagă au observat o tendinţă îngrijorătoare: toţi algoritmii lor comerciali care în trecut avuseseră succes păreau să nu mai funcţioneze, nereuşind sa se descurce nici măcar cu fondurile indexate. Cineva Ie bătea mereu pe propriul lor teren de joc.

Deşi numeroşi oameni puternici se opuneau valului schimbării, reacţia lor era absolut ineficientă, de parcă ar fi căzut intr-o capcană abil ticluită. Schimbări uriaşe aveau loc într-un ritm atât de ameţitor, încât era greu să le urmăreşti şi să vii cu un răspuns coordonat. în plus, nu era deloc limpede ce trebuia făcut. Dreapta politică tradiţională văzuse că-i sunt preluate majoritatea sloganurilor, dar reducerea impozitelor şi climatul de afaceri îmbunătăţit ajutau mai curând concurenţa de înaltă tehnologie. Practic toate întreprinderile tradiţionale cereau sprijin financiar, dar fondurile guvernamentale

erau limitate, aşa că se luptau fără speranţă între ele, în vreme ce mediile de informare Ie prezentau ca pe nişte dinozauri care cer subvenţii de la stat pur şi simplu din lipsă de competitivitate. Stânga politică tradiţională se opunea liberului schimb şi reducerii serviciilor sociale, dar era încântată de reducerea cheltuielilor militare şi de scăderea sărăciei. Protestele ei erau amuţite de incontestabilul fapt că serviciile sociale se îmbunătăţiseră odată ce erau oferite de companii idealiste, iar nu de stat. Sondajele de opinie arătau că majoritatea respondenţi- lor din lumea întreagă percepeau o îmbunătăţire a calităţii vieţii şi credeau că lucrurile se îndreaptă în general într-o direcţie bună. Explicaţia de ordin matematic era simplă: înainte de Prometeu, jumătatea mai săracă a planetei deţinea doar 4% din venitul global, ceea ce a permis companiilor controlate de omegi să câştige inimile (şi voturile) celor săraci, împărţind cu ei doar o mică parte din profituri.

CONSOLIDAREA

Ca urmare, partidele care îmbrăţişau cele şapte sloganuri ale ome- gilor au obţinut pretutindeni victorii răsunătoare în alegeri. în campanii atent optimizate, ele se declarau de centru, denunţând lăcomia dreptei care cerea subvenţii de Ia stat şi căuta să declanşeze conflicte, şi acuzând stânga că vrea să introducă mereu impozite. Nimeni n-a observat că Prometeu îşi alesese cu grijă candidaţi! şi trăgea sforile pentru a le asigura victoria.

înainte de Prometeu, luase avânt o mişcare în favoarea unui venit de bază universal, care propunea un venit minim pentru toată lumea, obţinut prin impozitare, ca remediu la şomajul tehnologic. Mişcarea a implodat atunci când au apărut proiectele comunitare ale

corporaţiilor, fiindcă imperiul de afaceri controlat de omegi oferea de fapt exact acelaşi lucru. Sub pretextul îmbunătăţirii coordonării proiectelor lor comunitare, un grup internaţional de companii a înfiinţat Alianţa Umanitară, o organizaţie nongu- vemamentală având drept scop identificarea şi subvenţionarea celor mai importante eforturi umanitare pretutindeni în lume. Curând, a ajuns s-o susţină practic întregul imperiu Omega, şi s-au lansat proiecte globale la o scară fără precedent, chiar şi în ţări ocolite de explozia tehnologică, făcând să crească nivelul educaţiei, asistenţei medicale, prosperităţii şi administraţiei. Evident, din culise Prometeu oferea planuri minuţios elaborate: în loc să împrăştie bani cu ţârâita, ca în propunerile pentru un venit de bază, Alianţa (aşa cum începuse să fie numită) îi privilegia pe cei care lucrau pentru cauza ei. Ca urmare, o mare parte din populaţia lumii era recunoscătoare şi fidelă Alianţei – deseori mai mult decât faţă de propriul guvern.

Cu trecerea timpului, Alianţa şi-a asumat un rol tot mai important în conducerea lumii, iar puterea guvernelor naţionale s-a erodat continuu. Bugetele naţionale scădeau prin reducerea impozitelor, în vreme de bugetul Alianţei depăşea cu mult toate bugetele naţionale însumate. Rolurile tradiţionale ale guvernelor au devenit redundante şi irelevante. Alianţa oferea de departe cele mai bune servicii sociale, educaţie şi infrastructură. Mediile de informare dezamorsaseră conflictele internaţionale până în punctul în care cheltuielile militare au devenit aproape inutile, iar creşterea prosperităţii eliminase majoritatea cauzelor vechilor conflicte, bazate pe lupta pentru resurse insuficiente. Câţiva dictatori s-au opus violent acestei noi ordini mondiale şi au refuzat să fie cumpăraţi, dar au fost cu toţii răsturnaţi prin lovituri de stat şi revolte atent orchestrate.

Omegii au încheiat cea mai spectaculoasă tranziţie din istoria vieţii pe Pământ. Planeta era pentru prima dată condusă de o unică putere, amplificată de o inteligenţă atât de vastă, încât putea permite vieţii să prospere vreme de miliarde de ani pe Pământ şi în cosmos – dar care erau de fapt planurile lor?

CAPITOLUL 1

Sunteţi invitaţi la cea mai importantă dezbatere a epocii noastre

Tehnologia oferă vieţii posibilitatea de a prospera ca niciodată până acum – sau de a se autodistruge.

Institutul pentru Viitorul Vieţii

La 13,8 miliarde de ani de la naşterea lui, universul nostru s-a trezit şi devine conştient de sine. De pe o mică planetă albastră, minuscule părţi conştiente ale universului au început să privească spre cosmos cu telescoape, descoperind în repetate rânduri că tot ce credeau că există e doar o infimă parte din ceva mult mai vast: un sistem solar, o galaxie şi un univers cu peste o sută de miliarde de alte galaxii aranjate într-o structură complicată de grupuri, roiuri şi superroiuri. Deşi aceşti privitori la stele con- ştienţi de sine se contrazic în multe privinţe, ei cad de regulă de acord asupra faptului că galaxiile sunt frumoase şi inspiră un sentiment înălţător.

Dar frumuseţea este în ochii privitorului, nu în legile fizicii, aşa încât, înainte ca universul nostru să se

trezească, nu exista nici o frumuseţe. Asta face şi mai miraculoasă deşteptarea noastră cosmică – ea a transformat universul dintr-un zombi fără conştiinţă de sine într-un ecosistem viu adăpostind autoreflecţia, frumosul şi speranţa, căutarea unor obiective şi a unei semnificaţii. Dacă nu s-ar fi deşteptat, universul ar fi rămas, cred eu, complet inutil, o uriaşă irosire de spaţiu. Dacă universul va adormi pe veci la loc, din pricina vreunei catastrofe cosmice sau a autodistrugerii, va deveni lipsit de sens.

Pe de alta parte, lucrurile ar putea merge şi mai bine. Deocamdată nu ştim dacă noi, oamenii, suntem singurii privitori la stele din univers, sau măcar primii, dar am aflat deja destule despre universul nostru pentru a şti că se poate deştepta încă mai mult decât a făcut-o până acum. Suntem pesemne la fel ca primul licăr al conştiinţei de sine care apare atunci când te trezeşti dimineaţa din somn: premoniţia unei conştiinţe mult maî vaste la care ajungi după ce deschizi ochii şi eşti pe deplin treaz. Poate că viaţa se va răspândi prin cosmos şi va prospera vreme de miliarde sau bilioane de ani – şi poate că asta se va datora hotărârilor pe care le luăm aici, pe mica noastră planetă, în cursul vieţii noastre.

0 SCURTĂ ISTORIE A COMPLEXITĂŢII

Cum a apărut aşadar această miraculoasă deşteptare? N-a fost un eveniment izolat, ci un pas dintr-un neîncetat proces durând 13,8 miliarde de ani, care face ca universul nostru să devină tot mai complex şi mai interesant – şi care continuă într-un ritm accelerat.

Ca fizician, mă consider norocos că am petrecut o bună parte din ultimul sfert de veac contribuind la

înţelegerea istoriei noastre cosmice, şi a fost o extraordinară aventură a cunoaşterii. Pe când absolveam facultatea ne întrebam dacă vârsta universului este de 10 sau 20 de miliarde de ani, acum ne întrebăm dacă vârsta lui este de 13,7 sau 13,8 miliarde de ani, iar asta graţie unor mai bune telescoapeşi calculatoare, şi unei mai bune înţelegeri. Nu ştim încă exact ce a declanşat big bang-ul sau dacă a fost într-adevăr începutul tuturor lucrurilor, ori doar urmarea unei etape anterioare. înţelegem totuşi destul de bine ce s-a întâmplat *după* big bang, graţie unei avalanşe de măsurători de înaltă precizie, aşa încât permi- teţi-mi să rezum cei 13,8 miliarde de ani ai istoriei cosmice.

La început a fost lumină. în prima fracţiune de secundă după big bang, întreaga parte a spaţiului pe care o pot în principiu observa telescoapele noastre („universul observabil" sau „universul nostru") era mult mai fierbinte şi mai strălucitoare decât miezul Soarelui nostru, şi se extindea rapid. Deşi sună spectaculos, era banal, fiindcă universul nostru nu conţinea decât o supă densă, fierbinte, plictisitor de uniformă şi lipsită de viaţă, alcătuită din particule elementare. Lucrurile arătau pretutindeni cam la fel, iar singura structură interesantă era dată de nişte slabe unde acustice aleatorii care făceau ca supa să fie cu vreo 0,001% mai densă în anumite locuri. Aceste unde slabe se crede că au apărut ca fluctuaţii cuantice – pentru că principiul incertitudinii al lui Heisenberg din mecanica cuantică nu permite nici unui lucru să fie cu desăvârşire plictisitor şi uniform.

Pe măsură ce universul nostru se extindea şi se răcea, devenea tot mai interesant, odată ce particulele lui se combinau în obiecte mai complexe. în cursul primei fracţiuni de secundă, forţa nucleară tare a adunat cuarcii în protoni (nuclee de hidrogen) şi

neutroni, iar în câteva minute unii dintre ei au fuzionat, formând nuclee de heliu. Peste vreo 400 000 de ani, forţa electromagnetică a combinat aceste nucleele cu electroni pentru a forma primii atomi. Pe măsură ce universul continua să se extindă, atomii se răceau treptat, alcătuind un gaz întunecat şi rece, iar întunericul acestei prime nopţi a durat circa 100 de milioane de ani. După lunga noapte s-au ivit zorii cosmosului nostru, când forţa gravitaţională a reuşit să amplifice acele fluctuaţii din gaz, atrăgând atomii laolaltă pentru a forma primele stele şi galaxii. Aceste prime stele au produs căldură şi lumină făcând să fuzioneze atomii de hidrogen în atomi mai grei cum sunt cei de carbon, oxigen şi siliciu. Când primele stele au murit, mulţi dintre atomii creaţi de ele au fost reciclaţi în cosmos, alcătuind planete ce se roteau în jurul stelelor din generaţia a doua.

La un moment dat, un grup de atomi s-a aranjat într-o structură complexă care se putea menţine şi replica. Curând, au apărut două exemplare identice, apoi numărul lor a continuat să se dubleze. Sunt de-ajuns 40 de dublări pentru a se ajunge la un bilion, aşa încât acest prim replicator a devenit repede o forţă care trebuia luată în considerare. Apăruse viaţa.

Problema definirii vieţii e larg dezbătută. Există o mulţime de definiţii, unele incluzând cerinţe extrem de precise, cum ar fi alcătuirea din celule, care ar putea descalifica atât viitoarele maşini inteligente, cât şi civilizaţiile extraterestre. Fiindcă nu vrem să ne limităm perspectiva asupra viitorului vieţii la speciile întâlnite până acum, vă propun să definim viaţa în sens foarte larg, ca pe un proces care îşi poate menţine şi replica propria-i complexitate. Nu materia (alcătuită din atomi) este replicată, ci informaţia (alcătuită din biţi) care precizează cum sunt aranjaţi atomii. Când o bacterie face o copie a ADN-ului ei, nu sunt creaţi noi atomi, ci un nou set de atomi e aranjat în acelaşi tipar ca originalul, copiindu-se astfel informaţia. Cu alte cuvinte, putem privi viaţa ca pe un sistem de prelucrare a informaţiei autoreplicator a cărui informaţie (software) îi determină deopotrivă comportamentul şi planurile pentru hardware.

La fel ca universul nostru, viaţa a devenit treptat tot mai complexă şi mai interesantă,[‡] iar după cum voi arăta, găsesc util să clasificăm formele de viaţă în trei niveluri de complexitate: Viaţa i.o, 2.0 şi 3.0. Aceste trei niveluri sunt prezentate schematic în figura 1.1.

Nu ştim deocamdată cum, când şi unde a apărut pentru prima dată viaţa în universul nostru, dar există dovezi convingătoare că aici, pe Pământ, viaţa a apărut cu vreo 4 miliarde de ani în urmă. Nu după multă vreme

‡ De ce a devenit viaţa tot mai complexă? Evoluţia răsplăteşte viaţa care e suficient de complexă pentru a prezice şi exploata regularităţile din mediul ei, aţa încât într-un mediu mai complex va evolua o viaţă mai complexă ţi mai inteligentă. Această viaţă mai deşteaptă creează un mediu mai complex pentru formele de viaţă aflate în competiţie, care, la rândul lor, evoluează devenind mai complexe, creând în cele din urmă ecosistemul unei vieţi extrem de complexe. (N. o.)

pe planeta noastră forfotea o mare bogăţie de forme de viaţă. Cele de succes, care curând şi-au surclasat competitorii, au putut cumva reacţiona la mediul lor. Mai precis, ele au fost ceea ce informaticienii numesc „agenţi inteli- genţi“: entităţi care culeg informaţii despre mediu de la senzori,

iar apoi procesează informaţiile pentru a hotărî cum să reacţioneze la mediu. Asta poate include o procesare foarte complexă a informaţiei, ca atunci când foloseşti informaţia primită de la ochi şi urechi pentru a hotărî ce să spui într-o conversaţie. Dar poate include şi un hardware şi software destul de simple.

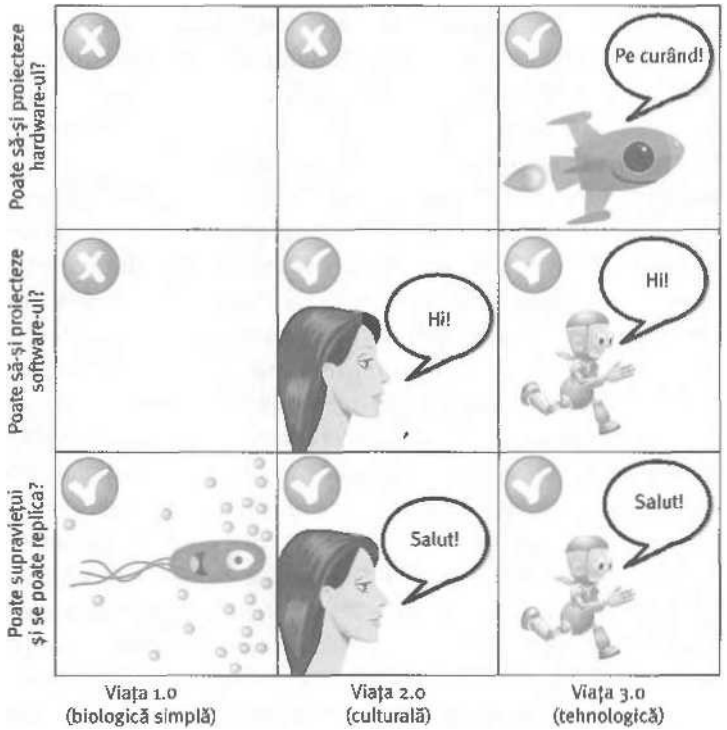


Figura 1.1 Cele trei etape ale vieţii: evoluţie biologică, evoluţie culturală şi evoluţie tehnologică. Viaţa 1.0 e incapabila să-şi reproiecteze hardware-ul sau software-ul în timpul vieţii: ambele sunt determinate de ADN-ul ei, şi se modifică doar prin evoluţia de-a lungul a numeroase generaţii. Dimpotrivă, Viaţa 2.0 poate reproiecta mult din software-ul ei: oamenii pot căpăta noi deprinderi complexe – de pildă, limbaje, sporturi sau profesiuni – şi-şi pot

actualiza perspectiva asupra lumii şi obiectivele. Viaţa 3.0, care nu există încă pe Pământ, îşi poate reproiecta spectaculos nu numai software-ul, dar şi hardware-ul, în loc să aştepte ca acesta să evolueze treptat de-a lungul generaţiilor.

De pildă, multe bacterii au un senzor care măsoară concentraţia de glucide din lichidul din jurul lor, şi pot înota folosind structuri propulsoare numite flageli. Hardware-ul care leagă senzorul de flageli poate pune în aplicare următorul algoritm simplu, dar util: „Dacă senzorul meu pentru concentraţia glucidelor raportează o valoare mai mică decât cu câteva secunde în urmă, atunci inversează rotaţia flagelilor mei, aşa încât să schimb direcţia.“

Aţi învăţat să vorbiţi şi nenumărate alte lucruri. Bacteriile însă nu prea se pricep la învăţat. ADN-ul lor le proiectează nu doar hardware-ul (senzorii pentru glucide, flagelii), ci şi software-ul. Ele nu învaţă niciodată să înoate spre glucide; în schimb, acel algoritm a fost codificat în ADN-ul lor de la început. A existat, desigur, un fel de proces de învăţare, dar el n-a avut loc în cursul vieţii unei anumite bacterii, ci s-a desfăşurat în evoluţia anterioară a acelei specii de bacterii, printr-un îndelungat şir de încercări şi eşecuri de-a lungul multor generaţii, în care selecţia naturală a favorizat acele mutaţii întâmplătoare din ADN care îmbunătăţeau consumul de glucide. Unele mutaţii au îmbunătăţit proiectarea flagelilor şi a altor componente ale hardware-ului, iar altele au îmbunătăţit sistemul de procesare a informaţiei care aplică algoritmul de găsire a glucidelor şi alte programe.

Asemenea bacterii sunt un exemplu pentru ceea ce numesc „Viaţa 1.0“: *viaţa în care hardware-ul ţi software-ul sunt rezultatul evoluţiei, nu sunt proiectate.* Prin software înţeleg toţi algoritmii şi toate cunoştinţele pe care le foloseşti pentru a procesa informaţia primită de la simţuri şi hotărăşti ce să faci – totul, de la capacitatea de a-ţi recunoaşte prietenii când îi vezi,

până la cea de a merge, citi, scrie, calcula, cânta şi spune bancuri.

Nu erai în stare să îndeplineşti nici una dintre aceste sarcini când te-ai născut, deci întreg acest software a fost programat în creierul tău mai târziu, prin procesul pe care îl numim învăţare. Deşi planul de învăţare din copilărie e în mare parte determinat de familie şi profesori, care hotărăsc ce trebuie să asimilezi, treptat capeţi tot mai multă putere pentru a-ţi proiecta software-ul. Poate că şcoala îţi permite să alegi o limbă străină: Vrei să instalezi în creierul tău un modul software care te ajută să vorbeşti engleza, sau unul care te ajută să vorbeşti franceza? Vrei să înveţi să joci tenis sau şah? Vrei să studiezi pentru a deveni avocat sau farmacist? Vrei să afli mai multe despre inteligenţa artificială (IA) şi viitorul vieţii citind o carte pe tema asta'

Capacitatea Vieţii 2.0 de a-şi proiecta software-ul îi permite să fie mult mai deşteaptă decât Viaţa 1.0.0 inteligenţă superioară presupune deopotrivă mult hardware (constând în atomi) şi mult software (constând în biţi). Faptul că majoritatea hardware-ului uman e adăugat după naştere (prin creştere) este avantajos, căci dimensiunea la care ajungi în cele din urmă nu e determinată de lărgimea colului uterin al mamei. De asemenea, faptul că majoritatea software-ului nostru uman e adăugat după naştere (prin învăţare) este avantajos, căci inteligenţa la care ajungem în cele din urmă nu e limitată de cantitatea de informaţie care ne poate fi transmisă în momentul conceperii prin ADN, în stilul Vieţii 1.0. Eu cântăresc de vreo douăzeci şi cinci de ori mai mult decât atunci când m-am născut, iar conexiunile sinaptice care leagă neuronii din creierul meu pot stoca de vreo sută de mii de ori mai multă informaţie decât ADN-ul cu care m-am născut. Sinapsele stochează toate cunoştinţele şi deprinderile tale însumând aproximativ 100 de terabyţi de informaţie, în

vreme ce ADN-ul stochează doar aproximativ un gigabyte, care abia ajunge pentru un singur film. Este deci fizic imposibil ca un prunc să se nască vorbind perfect limba maternă şi gata să dea examenul de admitere la liceu: informaţia nu poate fi pre-încărcată în creierul lui, fiindcă modulul de informaţie principal primit de la părinţi (ADN-ul lui) nu are capacitatea suficientă de stocare a informaţiei.

înzestrarea de a-şi proiecta software-ul îi permite Vieţii 2.0 să fie nu doar mai deşteaptă decât Viaţa 1.0, dar şi mai flexibilă. Dacă mediul se schimbă, 1.0 se poate adapta numai evoluând lent de-a lungul a numeroase generaţii. Viaţa 2.0 se poate însă adapta aproape instantaneu, prin actualizarea software-ului. De pildă, bacteriile care se confruntă frecvent cu antibiotice pot dezvolta rezistenţă la medicamente în cursul multor generaţii, dar o bacterie individuală nu-şi schimbă comportamentul; în schimb, o fată care află că are alergie la alune îşi va modifica imediat comportamentul, evitând alunele. Această flexibilitate oferă Vieţii 2.0 un avantaj încă şi mai mare la nivelul populaţiei: dacă informaţia din ADN-ul uman n-a evoluat spectaculos în ultimii 50 000 de ani, informaţia stocată colectiv în creierele noastre, în cărţi şi calculatoare a explodat. Prin instalarea unui modul software care ne permite să comunicăm prin limbajul verbal sofisticat, am făcut astfel încât majoritatea informaţiei utile stocată în creierul unei persoane să poată fi copiată în alte creiere, supravieţuind morţii creierului originar. Prin instalarea unui modul software care ne permite să citim şi să scriem, am putut stoca şi deţine în comun mult mai multă informaţie decât sunt în stare să memoreze oamenii. Prin dezvoltarea unui software cerebral capabil să producă tehnologie (studiind ştiinţele şi ingineria), am făcut ca mare parte din informaţie să poate fi accesată de aproape oricine cu doar câteva clicuri.

Această flexibilitate a permis Vieţii 2.0 să domine

Pământul. Eliberate de cătuşele geneticii, cunoştinţele însumate ale omenirii au sporit accelerat, fiecare realizare înlesnind-o pe următoarea: limbajul, scrierea, tiparul, ştiinţa modernă, calculatoarele, internetül etc. Această evoluţie culturală tot mai rapidă a softwa- reului nostru comun a devenit forţa dominantă care ne făureşte viitorul, făcând ca lenta evoluţie biologică să devină irelevantă.

în ciuda celor mai puternice tehnologii de care dispunem astăzi, toate formele de viaţă cunoscute rămân însă fundamental limitate de hardware-ul lor biologic. Nici una nu poate trăi un milion de ani, nu poate memora întreaga Wikipedie, înţelege întreaga ştiinţă la care am ajuns şi nu poate zbura în spaţiu fără o navă. Nici una nu poate transforma cosmosul nostru în mare parte lipsit de viaţă într-o biosferă diversă care va înflori miliarde sau bilioane de ani, permiţând universului să-şi împlinească în sfârşit posibilităţile şi să se deştepte pe deplin. Pentru toate acestea e nevoie ca viaţa să facă încă un pas înainte, către Viaţa 3.0, care poate proiecta nu doar software-ul, dar şi hardware-ul ei. Cu alte cuvinte, Viaţa 3.0 e stăpâna propriului ei destin, eliberată în cele din urmă complet de cătuşele evoluţiei.

Graniţele dintre cele trei etape ale vieţii nu sunt precis trasate. Dacă bacteriile reprezintă Viaţa 1.0, iar oamenii Viaţa 2.0, atunci pe şoareci i-am putea clasifica drept I.I: ei pot învăţa o sumedenie de lucruri, dar nu suficiente pentru a dezvolta limbajul sau a inventa intemetul. în plus, fiindcă le lipseşte limbajul, ceea ce învaţă se pierde atunci când mor, nu e transmis generaţiei următoare. în mod asemănător, am putea spune că oamenii de azi ar trebui socotiţi Viaţa 2.1: putem realiza mici upgradări ale hardware-ului, cum sunt implantarea dinţilor şi genunchilor artificiali, stimulatoarelor cardiace, dar nu şi ceva atât de spectaculos cum ar fi să devenim de zece ori mai înalţi sau să

dobândim un creier de o mie de ori mai mare.

Pe scurt, putem împărţi dezvoltarea vieţii în trei etape, diferenţiate prin capacitatea vieţii de a se proiecta pe sine însăşi:

- Viaţa 1.0 (etapa biologică): îşi evoluează hardware-ul şi software-ul
- Viaţa 2.0 (etapa culturală): îşi evoluează hardware-ul, îşi proiectează mult din software
- Viaţa 3.0 (etapa tehnologică): îşi proiectează hardware-ul şi software-ul

După 13,8 miliarde de ani de evoluţie cosmică, dezvoltarea s-a accelerat spectaculos aici, pe Pământ: Viaţa 1.0 a apărut acum vreo patru miliarde de ani, Viaţa 2.0 (noi, oamenii) a apărut cu vreo sută de mii de ani în urmă, iar mulţi cercetători IA cred că Viaţa 3.0 ar putea apărea în cursul următorului secol, eventual chiar în timpul vieţii noastre, adusă de progresele IA. Ce anume se va întâmpla şi ce va însemna asta pentru noi? Acesta e subiectul cărţii de faţă.

CONTROVERSE

întrebările de mai sus iscă numeroase controverse, specialiştii în IA contrazicându-se aprig nu numai în predicţiile lor, dar şi în reacţiile lor afective, care variază de la optimism încrezător la îngrijorare acută. Nu există consens nici măcar în privinţa efectelor economice, juridice şi militare pe termen scurt, iar dezacordul creşte atunci când extindem orizontul de timp şi ne punem problema *inteligenţei artificiale generale* (IAG) – în special în legătură cu capacitatea IAG de a atinge nivelul uman şi de a-l depăşi, permiţând apariţia Vieţii 3.0. *Inteligenţa generală* poate îndeplini practic orice obiectiv, inclusiv învăţarea, spre deosebire

de inteligenţa limitată a unui program pentru jocul de şah.

Controversa privind Viaţa 3.0 se concentrează în jurul a două întrebări: când şi ce? Când se va întâmpla (dacă se va întâmpla) şi ce va însemna asta pentru omenire? Cred că există trei şcoli de gândire care trebuie luate în serios, pentru că fiecare include ex ■ perţi de nivel mondial. După cum se vede în figura 1.2, îi numesc *utopişti digitali, tehno-sceptici* şi *membri ai mişcării IA benefică.* Daţi-mi voie să vă prezint pe unii dintre cei mai elocvenţi exponenţi ai lor.

UTOPIŞTI DIGITALI

Pe când eram puşti mi-i închipuiam pe miliardari plini de morgă şi aroganţă. Când l-am întâlnit pentru prima oară la Google în 2008, Larry Page mi-a spulberat complet aceste clişee. îmbrăcat în jeanşi şi cu un tricou modest, şi-ar fi găsit locul Ia un picnic de la MIT (Massachusetts Institute of Technology – Institutul Tehnologic din Massachusetts). Vocea lui domoală şi zâmbetul prietenos m-au făcut să mă simt relaxat. Pe 18 iulie 2015 ne am întâlnit în Napa Valley la o petrecere dată de Elon Musk şi soţia sa de pe atunci, Talulah, şi am început o discuţie despre pornirile scatologice ale copiilor noştri. I-am recomandat cartea clasică în domeniu, *The Day Mp Butt Went Psycho* de Andy Griffith, iar Larry a comandat-o pe loc. Aproape uitasem că el ar putea intra în istorie drept cel mai influent om din toate timpurile: dacă viaţa digitală superinteligentă va apărea în timpul vieţii mele, cred că asta se va datora deciziilor lui Larry.

Am cinat apoi împreună cu soţiile noastre, Lucy şi Meia, vorbind despre caracterul inevitabil al dobândirii

conştiinţei de către maşini – pentru el, o falsă problemă. Mai târziu, după cocteiluri, el şi Elon au discutat însufleţit despre viitorul IA şi ce trebuie făcut.

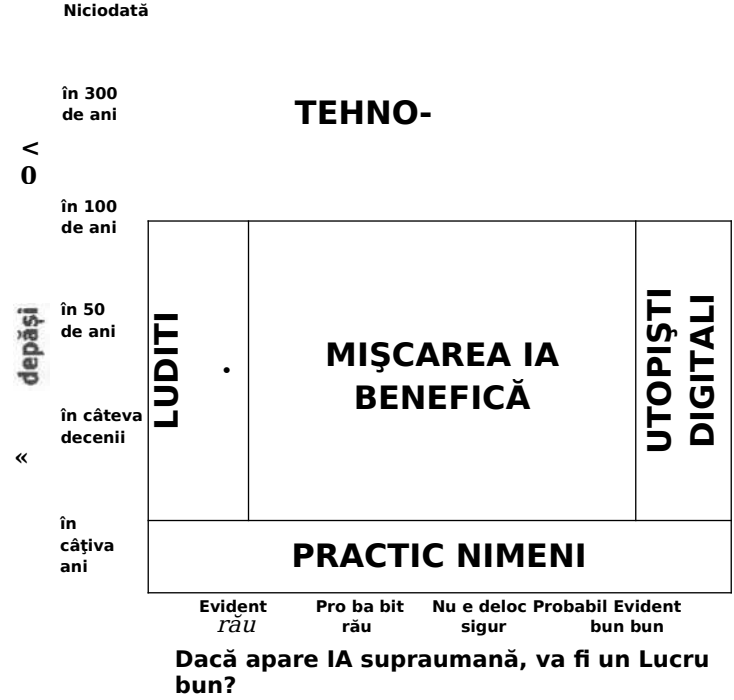


Figura 1.2 Majoritatea controverselor legate de inteligenţa artificială (dacă poate egala omul în orice sarcină cognitivă) se concentrează în jurul a două probleme: când se va întâmpla (dacă se va întâmpla) şi dacă va fi un lucru bun pentru omenire. Tehno-scepticii şi utopiştii digitali sunt de acord că n-ar trebui să ne batem capul cu ele, dar din motive diferite: primii sunt convinşi că inteligenţa artificială generală (IAG) la nivel uman nu va apărea în viitorul previzibil, în vreme ce ultimii cred cava apărea şi va fi fără îndoială un lucru bun. Mişcarea IA benefică susţine că îngrijorarea e îndreptăţită şi necesară, fiindcă cercetările şi dezbaterile privind siguranţa IA sporesc acum şansele unui rezultat bun. Ludiţii[§] sunt convinşi că rezultatul va fi rău şi se opun inteligenţei artificiale. Figura e parţial inspirată de Tim Urban.[1]

§ Ludiţii *(Iuddites)* au fost membrii unei organizaţii secrete a textiliştilor din Anglia secolului XIX, care distrugeau maşinile de ţesut, temându-se că munca lor manuală va fi înlocuită de acestea. în sens general, e vorba despre cei care, simţindu-se ameninţaţi, se opun progresului tehnologic. *(N.* t)

Spre dimineaţă, cercul spectatorilor şi chibiţilor continua să se lărgească, Larry a pledat cu pasiune pentru ceea ce eu numesc *utopism digital:* viaţa digitală este următorul pas natural şi dezira- bil în evoluţia cosmică, iar dacă noi vom lăsa libere minţile digitale, în loc să Ie oprim sau să le înrobim, rezultatul va fi aproape cu certitudine bun. ÎI consider pe Larry cel mai influent exponent al utopismului digital. Dacă viaţa se va răspândi vreodată în galaxia noastră şi dincolo de ea, susţinea el, atunci ar trebui s-o facă în formă digitală. Era îngrijorat că paranoia declanşată de IA ar întârzia utopia digitală şi/sau ar face ca militarii să preia contro- lul asupra IA, în contradicţie cu sloganul Google „Nu fiţi răi", Elon replica rugându-I pe Larry să lămurească detaliile argumentelor lui, întrebându-l, de pildă, de ce era convins că viaţa digitală nu va distruge toate lucrurile de care suntem ataşaţi. Larry îl acuza pe Elon că e „specist" – că tratează drept inferioare anumite forme de viaţă numai fiindcă se bazează pe siliciu, şi nu pe carbon. în capitolul 4 vom reveni asupra acestor aspecte interesante.

Deşi în acea noapte călduroasă de vară cei mai mulţi îl combă- teau pe Larry, utopismul digital apărat atât de elocvent de el are numeroşi susţinători de frunte. Roboticianul şi futurologul Hans Moravec a inspirat o întreagă generaţie de utopişti digitali cu clasica sa carte din 1988 *Mind Children,* tradiţie continuată şi amplifi- cată de inventatorul Ray Kurzweil. Richard Sutton, unul dintre pionierii domeniului IA numit „învăţare prin consolidare" *(reinforcement leaming),* a pledat cu pasiune în favoarea utopismului digital la o conferinţă din Puerto Rico despre care voi vorbi în curând.

Tehno-sceptici

Gânditorii aparţinând unui alt grup important nu-şi fac

nici ei griji în privinţa IA, dar dintr-un motiv diferit: crearea IAG supraumane li se pare atât de greu de înfăptuit în următoarele sute de ani, încât e ridicol să te preocupe acum. Această perspectivă tehno-scep- tică e limpede exprimată de Andrew Ng: „Să te temi de roboţi ucigaşi e ca şi cum te-ai teme de suprapopularea planetei Marte.“ Andrew a fost director de cercetare la Baidu, acel Google chinezesc, şi şi-a reluat argumentele când ne am întâlnit la o conferinţă la Boston. Mi-a spus şi că îngrijorarea legata de riscul IA ar putea fi o diversiune primejdioasă, încetinind progresele IA. Idei asemănătoare sunt împărtăşite şi de Rodney Brooks, fost profesor la MIT, creatorul aspiratorului robotic Roomba şi al robotului industrial Baxter. E interesant de observat că deşi utopiştii digitali şi tehno-scepticii cad de acord asupra faptului că IA nu trebuie să ne îngrijoreze, altminteri au puţine în comun. Majoritatea utopiştilor cred că IAG la nivel uman ar putea apărea într-un interval de timp între douăzeci şi o sută de ani, ceea ce pentru tehno-scep- tici sunt vise deşarte, exaltări ale computeriştilor. Când l-am întâlnit pe Rodney Brooks la o petrecere în decembrie 2014, mi-a spus că era 100% convins că nu se va întâmpla în timpul vieţii mele. „Sunteţi sigur că nu vreţi sa spuneţi 99%?“ l-am întrebat într-un e-mail, la care mi-a răspuns: „Nu 99%, 100%. Pur şi simplu n-o să se întâmple.”

Mişcarea IA benefică

Când l-am întâlnit pentru prima oară pe Stuart Russell, în iunie 2014 într-o cafenea pariziană, mi-a apărut ca exponentul tipic al unui gentleman britanic. Elocvent, grav, vorbind calm, dar cu o lucire cutezătoare în ochi, era întruchiparea modernă a lui Phileas Fogg, eroul copilăriei mele din romanul lui Jules Veme *Ocolul Pământului în optzeci de zile.* Deşi era unul dintre cei

mai celebri cercetători IA, coautor al cărţii de referinţă în domeniu, modestia şi căldura sa m-au făcut să mă simt în largul meu. Mi-a spus că progresele IA îl convinseseră că IAG la nivel uman poate fi obţinută în acest secol, dar că un rezultat benefic nu putea fi garantat. Trebuia mai întâi să răspundem la întrebări esenţiale, iar ele sunt atât de dificile, încât ar trebui să le cercetăm încă de-acum pentru a avea răspunsurile când va fi nevoie de ele.

Ideile lui Stuart sunt astăzi predominante, iar multe grupuri din lumea întreagă se ocupă cu studiul siguranţei IA pentru care pledează el. însă lucrurile n-au stat întotdeauna aşa. într-un articol din *Washington Post se* spunea că 2015 a fost anul în care siguranţa IA a intrat în centrul atenţiei. înainte, dacă vorbeai despre riscurile IA, erai tratat drept ludit care se opune progresului IA. După cum vom vedea în capitolul 5, temeri asemănătoare celor ale lui Stuart au fost exprimate pentru prima dată cu mai bine de o jumătate de secol în urmă de Alan Turing, părintele calculatorului, şi matematicianul Irving J. Good, care a lucrat cu Turing la spargerea codurilor germane în timpul celui de-al Doilea Război Mondial. în deceniul precedent, asemenea preocupări aveau doar câţiva gânditori independenţi care nu erau specialişti în IA – de pildă, Eliezer Yudkowsky, Michael Vassar şi Nick Bostrom fără ecou în rândul specialiştilor, care se concentrau asupra sarcinilor curente de îmbunătăţire a sistemelor IA. Dintre cercetătorii IA care nutreau anumite îngrijorări, mulţi ezitau să le dea glas de teamă că vor fi catalogaţi drept tehnofobi alarmişti.

Simţeam că această polarizare trebuia depăşită, aşa încât întreaga comunitate IA să participe la dezbaterea privind felul în care poate fi construită IA benefică. Din fericire, nu eram singurul, în primăvara lui 2014 am

înfiinţat o organizaţie non-profit numită Institutul pentru Viitorul Vieţii (Future of Life Institute, FLI; http://futureoflife.org) împreună cu soţia mea Meia, prietenul meu fizician Anthony Agguire, Viktoriya Krakovna, absolventă de Harvard, şi Jaan Tallinn, fondatorul Skype-ului. Scopul nostru era simplu: să contribuim la asigurarea viitorului vieţii, un viitor cât mai bun cu putinţă. Mai exact, simţeam că tehnologia dădea vieţii puterea fie de a cunoaşte o înflorire fără precedent, fie de a se autodistruge, iar noi preferam prima variantă.

A avut Ioc mai întâi o şedinţă de *brainstorming* la noi acasă, pe 15 martie 2014, la care au participat vreo treizeci de studenţi, profesori şi persoane interesate de subiect din zona Bostonului. Am căzut de acord asupra faptului că, deşi trebuie să ne preocupe biotehnologiile, armele nucleare şi modificarea climei, principalul nostru obiectiv este aducerea în prim-plan a cercetărilor privind siguranţa IA. Colegul meu de la MIT Franck Wilczek, care a primit Premiul Nobel pentru contribuţii Ia fizica cuarcilor, a propus să începem prin a scrie un editorial de opinie în care să atragem atenţia asupra temei, făcând-o mai greu de ignorat. Am luat legătura cu Stuart Russell (pe care nu-l întâlnisem încă) şi Stephen Hawking, iar amândoi au fost de acord să fie coautori, alături de mine şi de Franck. Editorialul ne-a fost respins de *The New York Timesşi* de multe alte ziare americane, aşa că l-am postat pe contul meu de blog de Ia *Huffington Post* Spre bucuria mea, Ari- anna Huffington însăşi mi-a trimis un mesaj în care spunea „Sunt încântată să-l avem! îl vom posta la #1”, iar această plasare la începutul paginii a declanşat un val de acoperire media a siguranţei IA care a durat tot anul, cu intervenţiile lui Elon Musk, BilI Gates şi ale altor figuri proeminente din tehnologie. Cartea Iui Nick

Bostrom *Superintelligence* a apărut în acea toamnă, înteţind dezbaterea publică.

Următorul obiectiv al campaniei noastre pentru IA benefică a fost să-i reunim pe principalii cercetători IA într-o conferinţă la care să se lămurească neînţelegerile, să se ajungă la un consens şi să se avanseze planuri constructive. Ştiam că e greu de convins un asemenea grup de persoane ilustre să vină la o conferinţă organizată de outsideri pe care nu-i cunoşteau, mai ales că tema era controversată, aşa că ne-am străduit pe cât am putut – am interzis accesul presei, am ales ca loc de desfăşurare o staţiune din Puerto Rico, n-am cerut taxe de participare (graţie generozităţii lui Jaan Tallinn) şi am găsit cel mai nealarmist titlu imaginabil: „Viitorul IA: Posibilităţi şi provocări". în plus, ne-am asociat cu Stuart Russell, mulţumită căruia am lărgit comitetul de organizare pentru a include specialişti IA din mediul academic şi din industrie – între care, Demis Hassabis de la DeepMind, Google, care a arătat că IA îi poate învinge pe oameni chiar şi în jocuri precum go. Cunoscându-l pe Demis, mi-am dat seama că avea ambiţia de a face IA nu doar puternică, dar şi benefică.

Urmarea a fost o remarcabilă întrunire a unor mari minţi (figura 1.3). Cercetătorii IA s-au întâlnit cu economişti şi jurişti de vârf, lideri din tehnologie (între care Elon Musk) şi alţi gânditori

Figura 1.3 Conferinţa din ianuarie 2013 din Puerto Rico a reunit un grup remarcabil de cercetători IA şi din domeniile conexe. Pe rândul

din spate, de la stânga la dreapta: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Areas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Alţii, stând în picioare, de la stânga la dreapta: Marilyn Thompson, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, MaxTegmark, Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Viadeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krâmar, Geoff Anders, Vernor Vinge, Anthony Aguirre. în rândul din faţă: Sam Harris, Tomaso Poggio, Marin SoljaEif, Viktoriya Krakovna, Meia Chita-Tegmark.în spatele aparatului de fotografiat: Anthony Aguirre (introdus în imagine prin photoshop).

(între care Vemon Vinge, cel care a introdus termenul „singularitate”, tratat în capitolul 4). Rezultatul a depăşit cele mai optimiste aşteptări. Poate că a fost o combinaţie între vin şi soare, sau poate că pur şi simplu sosise momentul potrivit: în ciuda subiectului controversat, s-a ajuns la un consens remarcabil, care a fost

Figura 1.4 Deşi s-a spus adesea că Elon Musk se află în dispută cu comunitatea IA, s-a ajuns la un consens larg privind necesitatea cercetărilor de siguranţă IA. în fotografia făcută pe 4 ianuarie 2015, Tom Dietterich, preşedintele Asociaţiei pentru Progresul Inteligenţei Artificiale împărtăşeşte cu Elon entuziasmul pentru noul program de cercetare în siguranţa IA pe care Elon tocmai promisese să-l finanţeze. în spatele lor se zăresc Meia-Cita Tegmark şi Vikto- riya Krakovna, fondatoare ale FLI.

exprimat sub forma unei scrisori deschise[2] semnate în cele din urmă de peste opt mii de oameni, incluzând un veritabil „who's who" în IA. în esenţă, scrisoarea spunea că IA trebuia îndrumată pentru a deveni benefică, şi menţiona în detaliu temele de cercetare care urmau să fie abordate în vederea acestui obiectiv. Mişcarea pentru IA benefică intrase în prim-plan. Evoluţia ei ulterioară va fi prezentată în cartea de faţă.

Altă concluzie a conferinţei a fost că problemele ridicate de succesul LA nu sunt doar fascinante în plan intelectual, ci au şi o importanţă crucială în plan moral, fiindcă alegerile noastre pot afecta viitorul vieţii. Alegerile omenirii din trecut au avut uneori o mare însemnătate morală, dar ea a fost întotdeauna limitată: ne am revenit după dezastre, iar până şi imperiile cele mai falnice s-au prăbuşit în cele din urmă. Generaţiile din trecut ştiau că, aşa cum soarele va răsări a doua zi, şi oamenii vor avea a doua zi de înfruntat nenorociri

precum sărăcia, bolile şi războiul. Unii conferenţiari din Puerto Rico au susţinut însă că de data asta lucrurile

ar putea sta altfel: pentru prima oară am putea crea o tehnologie suficient de puternică pentru a pune capăt definitiv acestor nenorociri – sau pentru a pune capăt omenirii înseşi. Am putea făuri societăţi care să prospere ca niciodată, sau un stat global kafkian care supraveghează totul, atât de puternic încât să fie imposibil de răsturnat.

CONFUZII

Când am plecat din Puerto Rico eram convins că discuţiile despre viitorul IA trebuie să continue, fiindcă e cea mai importantă dezbatere a epocii noastre.** Este dezbaterea despre viitorul colectiv al nostru, al tuturor, deci n-ar trebui să se limiteze la cercetătorii IA Am scris cartea de faţă în speranţa că tu, dragă cititorule, vei participa la această dezbatere. Ce fel de viitor îţi doreşti? Vrei să construim arme autonome letale? Ce-ai vrea să se întâmple cu automatizarea muncii? Ce carieră i-ai sfătui să urmeze pe copiii de azi? Preferi ca noi joburi să le înlocuiască pe cele vechi, sau o societate fără joburi, în care toată lumea să se relaxeze, profitând de bogăţia produsă de maşini? Privind mai departe, ai vrea să creăm Viaţa 3.0 şi s-o răspândim în univers? Maşinile inteligente ne vor înlocui, vor coexista cu noi sau se vor contopi cu noi? Ce va însemna să fii om, în

** Dezbaterea despre IA e importantă atât pentru urgenţa, cât şi pentru impactul ei. în comparaţie cu modificarea climei, care va avea efecte dăunătoare peste cincizeci sau peste două sute de ani, mulţi experţi cred că IA va avea un impact mai puternic în următoarele decenii – şi ne-ar putea oferi tehnologia pentru a stăvili modificarea climei. în comparaţie cu problemele legate de război, terorism, şomaj, sărăcie, migraţie şi nedreptatea socială, apariţia IA va avea un impact global mai mare – iar în cartea de faţă vom vedea ce efecte, bune sau rele, ar putea avea asupra tuturor acestor probleme. *(N. a.)*

epoca inteligenţei artificiale? Ce ai vrea tu să însemne şi cum putem face ca viitorul să arate aşa cum îţi doreşti?

Scopul cărţii de faţă este de a te invita să participi la dezbatere. După cum am spus, sunt dezbateri fascinante în care mari experţi se află pe poziţii opuse, dar am văzut şi numeroase exemple de

Rezumat al terminologiei:

Viaţă	**Proces care îţi păstrează complexitatea ţi se replichează**
Viaţa 1.0	**Viaţa care îţi evoluează hardware-ul ţi software-ul (etapa biologică):**
Viaţa 2.0	**Viaţa care îţi evoluează hardware-ul, dar îţi proiectează mult din software (etapa culturală)**
Viaţa 3.0	**Viaţa care îţi proiectează hardware-ul ţi software-ul (etapa tehnologică)**
Inteligenţă	**Capacitatea de a îndeplini obiective complexe**
Inte.i«m artificială (1Ä)	**Inteligenţa ne biologi că**
Inteligenţă limitată	**Capacitatea de a îndeplini un set limitat de sarcini (deex., de a juca ţah, de a conduce 0 maşină)**
Inteligenţă generală	**Capacitatea de a îndeplini practic orice obiectiv, inclusiv învăţarea**
Inteligenţă universală	**Capacitatea de a dobândi inteligenţă generală, având acces la date sî resurse**
Inteligenţă artificială generală [la nivel uman] (IAG)	**Capacitatea de a îndeplini orice sarcină cognitivă, cel puţin la fel de bine ca oamenii**
IA la nivel uman	**IAG**
IA tare	**IAG**
Superinteligenţă	**Inteligenţă generală mult dincolo de nivelul uman**
Civilizaţie	**Grup în care mteracţionează forme inteligente de viaţă**
Conştiinţă	**Experienţă subiectivă**
Qualia	**Expresii individuale ale experienţei subiective**
Etică	**Principii care ne spun cum trebuie să ne comportăm**
Teleologie	**Explicarea lucrurilor pe baza obiectivelor sau scopurilor lor, nu pe baza cauzelor**
Comportament orientat spre obiective	**Comportament mai uţorde explicat prin efectele lui decât prin cauzele lui**
A avea un obiectiv	**A manifesta un comportament orientat spre obiective**
A av— un scop	**A sluji obiectivele proprii sau ale altei entităţi**
IA prietenoasă	**Superinteligenţă ale cărei obiecte sunt aliniate cu ale noastre**
Cyborg	**Hibrid om-masină**
Explozie a inteligenţei	**Auto perfecţiona re recurslvă rapidă conducând spre superinteligenţă**
Singularitate	**Explozie a inteligenţei**
Univers	**Regiunea spaţiului din care lumina a avuttimp să ajungă la noi în cursul celor 13,8 miliarde de ani scurţi de la big bang**

Tabelul 1.1 Numeroase confuzii legate de IA sunt provocate de sensurile diferite pe care le dăm termenilor, lată ce sens le dau eu în cartea de faţă. (Unele definiţii vor fr explicitate în capitolele următoare.) pseudodezbateri confuze plictisitoare. Pentru

a ne concentra asupra discuţiilor interesante şi întrebărilor deschise, iar nu asupra confuziilor, e cazul să lămurim cele mai frecvente neînţelegeri.

Există multe definiţii pentru termeni precum „viaţă“, „inteligenţă” şi „conştiinţă“, iar multe neînţelegeri sunt provocate de faptul că oamenii folosesc un cuvânt în sensuri diferite. Pentru a nu cădea în această capcană, am notat în tabelul r.i sensul în care folosesc eu termenii-cheie. Unele definiţii vor fi explicitate în capitolele următoare. Nu pretind că definiţiile mele sunt mai bune decât ale altora, vreau doar să risipesc confuziile. Observaţi că în general folosesc definiţii cuprinzătoare, care evită antropocentrismul şi pot fi aplicate maşinilor la fel de bine ca oamenilor. Vă invit să citiţi acum tabelul şi să reveniţi la el dacă sunteţi derutaţi de sensul în care folosesc termenii, în special în capitolele 4-8.

în plus faţă de confuzia provocată de terminologie, multe dezbateri despre IA sunt afectate de neînţelegeri simple. Vă propun să le lămurim pe cele mai frecvente.

Miturile cronologiei

Primele neînţelegeri se referă la cronologia din figura 1.2: cât va dura până când maşinile vor depăşi cu mult IAG la nivel uman? Ideea că am cunoaşte răspunsul cu certitudine este greşită.

Un mit larg răspândit spune că vom ajunge Ia IAG supraumană în acest secol. De fapt, istoria e plină de previziuni exagerate legate de tehnologie. Unde sunt acele centrale de fuziune sau maşini zburătoare care ni s-au promis? Şi în privinţa IA s-a exagerat în trecut, au facut-o chiar unii dintre fondatorii domeniului. De pildă, John McCarthy (cel care a creat termenul „inteligenţă artificială”), Marvin Minsky, Nathaniel Rochester şi Claude Shannon au prevăzut că aceasta va fi obţinută în două luni cu calculatoare primitive: „Propunem ca un

studiu asupra inteligenţei artificiale să fie efectuat timp de două luni de către zece oameni în vara lui 1956 la Colegiul Dartmouth. [...] Se va încerca să se afle cum pot fi făcute maşinile să folosească limbajul, să formeze abstracţiuni şi concepte, să rezolve probleme rezervate acum oamenilor şi să

Mit

în 2100 superinteligența va fi

Mit

în 2100 superinteligența va fi

i-api

Poate apărea peste decenii, secole sau niciodată – experții IA se contrazic, iar noi pur și simplu nu

Mit

Numai lud iți I sunt îngrijorați de IA.

Fapt

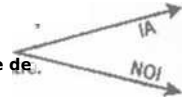
Mul(ercetători IA de vârf sunt preocupați.

îngrijorare

IA devine

îngrijorare

IA devine competentă, cu obiective diferite de



îngrijorare

IA devine

Mit

Roboții sunt principala

Fapt

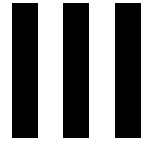
IA orientată diferit e principala problemă: nu are nevoie de corp, ci doar de conexiune

Mit

IA nu poate controla

Fapt

Inteligenta permite co itrolul: noi controlăm tigrii pentru că suntem



Mit

Mașinile nu pot avea

Fapt

Rachetele termoghidate au un obiectiv.

îngrijorare

Superinteligenț a va sosi

PANICA

I

îngrijorare

Va sosi cel puțin peste decenii, dar ar putea dura mult s-o facem

PLANIFIC Ă DIN

Figura 1.5 Mituri frecvent întâlnite privind IAsuperinteligentă.

se autoperfecţioneze. Credem că într-una sau mai multe dintre aceste probleme se poate obţine un progres important dacă un grup de oameni de ştiinţă atent selectaţi lucrează timp de o vară."

Pe de altă parte, conform unui contra-mit răspândit, ştim că *nu vom* ajunge la IAG supraumană în acest secol. Cercetătorii au făcut numeroase estimări privind timpul care ne separă de IAG supraumană, dar nu putem spune că probabilitatea pentru acest secol e zero, dat fiind jalnicul palmares al unor asemenea predic- ţii tehno-sceptice. De pildă, Emest Rutherford, pesemne cel mai mare fizician nuclearist al epocii sale, a spus în 1933 – chiar înainte ca Leo Szilard să găsească lanţul de reacţii nucleare – că energia nucleară e o prostie, iar în 1956 astronomul englez Richard Wooley a spus că discuţiile despre călătoria spaţială sunt „vorbe în vânt". Acest mit ia forma extremă în afirmaţia că nu se va ajunge niciodată la IAG supraumană, fiindcă e fizic imposibilă. Fizicienii ştiu totuşi că un creier e alcătuit din cuarci şi electroni aranjaţi pentru a acţiona ca un calculator puternic, şi că nici o lege a fizicii nu-i împiedică să se constituie într-o structură şi mai inteligentă.

S-au făcut câteva sondaje printre cercetătorii IA, cu întrebarea când vom avea IAG la nivel uman cu o probabilitate de cel puţin 50%, iar în nici un sondaj experţii n-au căzut de acord, aşa că pur şi simplu nu ştim. De pildă, intr-un asemenea sondaj făcut la con- ferinţa din Puerto Rico, răspunsul mediu a fost anul 2055, dar după unii cercetători va dura sute de ani sau mai mult.

Există şi un mit care spune că cei preocupaţi de IA cred că aceasta va sosi foarte curând, când de fapt majoritatea estimează că vor mai trece cel puţin câteva decenii. Ei susţin însă că, nefiind 100% *siguri că* nu va fi

în acest secol, e înţelept să începem încă de pe-acum cercetările legate de siguranţă. După cum vom vedea în cartea de faţă, multe probleme de siguranţă sunt atât de dificile, încât ar putea trece decenii până să fie rezolvate, şi ar trebui s-o facem înainte ca nişte programatori care beau Red Bull să se hotărască să treacă la IAG la nivel uman.

Miturile controversei

Altă idee falsă este că numai nişte ludiţi ignoranţi sunt preocupaţi de siguranţa IA şi pledează pentru cercetările în domeniu. Când Stuart Russell a menţionat asta în cuvântarea lui din Puerto Rico, publicul a izbucnit în râs. Se mai spune şi că sprijinul acordat cercetărilor în siguranţa IA e controversată. De fapt, pentru a face o investiţie modestă în cercetările privind siguranţa IA nu-i nevoie să fii convins că riscurile sunt mari, la fel cum o investiţie modestă în asigurarea locuinţei e justificată pur şi simplu de probabilitatea nenulă de a-ţi lua foc casa.

Cred că mediile de informare au făcut ca dezbaterea despre siguranţa IA să pară mai controversată decât e în realitate. în fond, teama se vinde bine, iar articolele care folosesc citate scoase din context pentru a proclama sfârşitul iminent primesc mai multe clicuri decât cele echilibrate şi nuanţate. Prin urmare, doi oameni care au aflat unul despre poziţia celuilalt doar din citate media vor crede probabil că se contrazic mai mult decât se întâmplă în realitate. De pildă un tehno-sceptic care află despre poziţia lui Bill Gates dintr-un tabloid britanic şi-ar putea închipui că pentru acesta superinteligenţa e iminentă. Apoi, un membru al mişcării IA benefică, necunoscând despre poziţia lui Andrew Ng altceva decât afirmaţia menţionată mai sus privind suprapopularea planetei Marte, ar putea crede că acestuia nu-i pasă de siguranţa IA Eu ştiu că-i pasă,

dar cum Andrew crede că la IA se va ajunge în- tr-un viitor îndepărtat, e interesat în primul rând de provocările IA pe termen scurt.

Mituri despre risc

Am citit exasperat acest titlu scris cu litere de-o şchioapă în *Daily Maik** „Stephen Hawking ne previne că ascensiunea roboţilor ar putea fi catastrofală pentru omenire.“ Mi-a fost dat să văd nenumărate asemenea articole. De regulă, sunt însoţite de imaginea unui robot malefic purtând o armă, şi sugerează că ar trebui să fim îngrijoraţi de roboţi care se răscoală şi ne ucid pentru că au devenit conştienţi şi/sau răi. Asemenea articole rezumă tocmai scenariul care pe colegii mei din IA nu îi preocupă. El combină trei elemente: îngrijorarea faţă de *conştienţă, rău* şi *roboţi*

Când conduci maşina ai o experienţă subiectivă a culorilor, sunetelor etc. Dar o maşină care se conduce singură are o asemenea experienţă subiectivă? Simte ceva, sau e doar un zombi inconştient, fără experienţe subiective? Deşi acest mister al conştiinţei e interesant în sine, şi vom vorbi despre el în capitolul 8, este irelevant pentru riscul IA. Dacă te calcă o maşină fără şofer, n-are importanţă pentru tine dacă ea se simte în mod subiectiv conştientă. La fel, ce ne afectează pe noi, oamenii, este ce *face,* nu cum se simte în mod subiectiv IA.

Teama de maşini care devin rele e o altă falsă problemă. Adevăratul motiv de îngrijorare nu e răutatea, ci competenţa. O IA superinteligentă se pricepe prin definiţie foarte bine să-şi atingă obiectivele, oricare ar fi ele, aşa că trebuie să ne asigurăm că obiectivele ei sunt aliniate cu ale noastre. Pesemne că nu eşti un om care urăşte furnicile şi le calcă din răutate, dar dacă proiectezi o centrală hidroelectrică, iar un muşuroi

de furnici se află în regiunea care va fi inundată, ghinionul lor. Mişcarea IA benefică urmăreşte să împiedice plasarea omenirii în locul acelor furnici.

Neînţelegerea legată de conştiinţă se leagă de mitul că maşinile nu pot avea obiective. Evident, maşinile pot avea obiective în sensul restrâns de a manifesta un comportament orientat spre obiective: comportamentul unei rachete termoghidate se explică cel mai simplu prin obiectivul de a lovi ţinta. Dacă te simţi ameninţat de o maşină ale cărei obiective diferă de ale tale, atunci tocmai aceste obiective în sens restrâns te neliniştesc, nu dacă maşina e conştientă şi percepe vreun scop. Dacă eşti fugărit de o rachetă termoghidată, probabil că nu vei exclama „Nu-mi pasă, pentru că maşinile nu pot avea obiective!”.

Sunt de partea lui Rodney Brooks şi a altor pionieri ai roboticii care se simt pe nedrept demonizaţi în tabloide fiindcă unii jurnalişti au făcut o fixaţie pentru roboţi şi-şi împodobesc articolele cu monştri de metal cu ochi roşii strălucitori. De fapt, principala preocupare a mişcării IA benefică nu sunt roboţii, ci inteligenţa însăşi – mai precis, inteligenţa ale cărei obiective diferă de ale noastre. Ea n-are nevoie de un corp robotic pentru a ne face rău, e suficientă o conexiune intemet. În capitolul 4 vom vedea cum aceasta poate păcăli pieţele financiare, poate întrece în iscusinţă cercetătorii umani, poate manipula lideri politici şi dezvolta arme pe care noi nici măcar nu le-am înţelege. Chiar dacă i-ar fi imposibil fizic să construiască roboţi, o IA superinteligentă şi extrem de bogată ar putea cu uşurinţă plăti sau manipula milioane de oameni pentru a-i îndeplini inconştienţi poruncile, ca în romanul SF al lui William Gibson *Neuromancer.*

Ideea greşită despre roboţi se leagă de mitul că maşinile nu pot controla oamenii. Inteligenţa face cu

putinţă controlul: oamenii controlează tigrii nu pentru că sunt mai puternici, ci pentru că sunt mai deştepţi. Asta înseamnă că, dacă încetăm să fim cei mai deştepţi de pe planetă, s-ar putea şi să pierdem controlul.

Figura 1.5 rezumă aceste idei false frecvent întâlnite, aşa că ne putem debarasa de ele, pentru a ne concentra asupra controverselor legitime deloc puţine, după cum vom vedea.

DRUMUL DE URMAT

în restul acestei cărţi vom explora împreună viitorul vieţii cu IA. Vă propun să navigăm organizat prin acest subiect complex, explorând mai întâi, conceptual şi cronologic, întreaga poveste a vieţii, iar apoi să ne ocupăm de obiectivele şi semnificaţia acţiunilor de urmat pentru a ne făuri viitorul dorit.

în capitolul *2* vom vorbi despre bazele inteligenţei şi felul în care materia neînsufleţită poate fi rearanjată pentru a memora, calcula şi învăţa. îndreptându-ne spre viitor, povestea noastră se ramifică în mai multe scenarii definite de răspunsurile la anumite întrebări-cheie. Figura 1.6 rezumă întrebările-cheie pe care le vom întâlni pe măsură ce înaintăm în timp spre o IA tot mai avansată.

încă de pe acum suntem confruntaţi cu alegerea de a începe sau nu o cursă a înarmărilor IA şi cu întrebarea ce putem face pentru ca sistemele IA de mâine să fie robuste şi lipsite de erori. Dacă impactul economic al IA continuă să crească, trebuie să hotărâm cum să ne modernizăm legile şi ce cariere să-i sfătuim pe copiii noştri să urmeze pentru a evita joburi care vor fi curând automatizate. Vom examina asemenea chestiuni cu

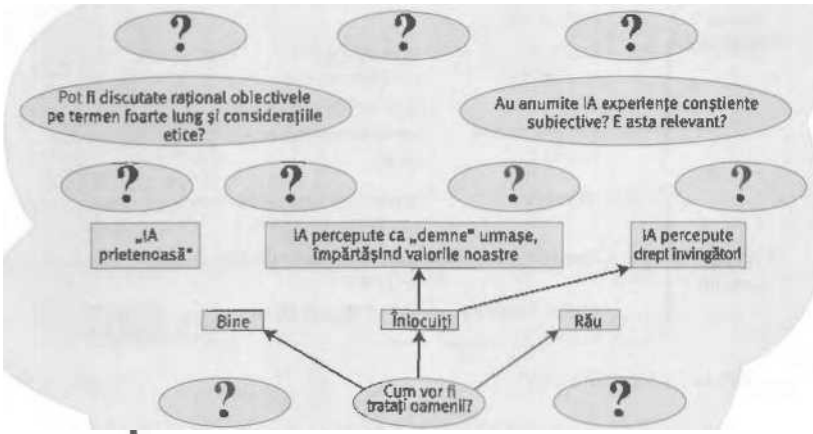
bătaie scurtă în capitolul 3.

Dacă IA avansează până la nivel uman, trebuie să ne întrebăm cum ne putem asigura că e benefică şi dacă e cazul să făurim o societate a relaxării care prosperă fără joburi. De aici se naşte întrebarea dacă o explozie a inteligenţei sau o lentă creştere constantă pot împinge IAG mult dincolo de nivelul uman. în capitolul 4 vom examina o gamă largă de asemenea scenarii, iar în capitolul 5 vom cerceta spectrul posibilităţilor care se deschid, de la cele distopice la cele utopice. Cine comandă – oamenii, IA sau cyborgii? Sunt oamenii trataţi bine sau rău? Suntem oare înlocu- îţi, iar în acest caz, îi percepem pe înlocuitori drept învingători sau demni urmaşi? Sunt curios să ştiu pe care dintre scenariile din capitolul 5 îl preferaţi! Am creat un site, http://AgeOfAi.org, unde vă puteţi spune părerea şi intra în dialog.

în fine, vom înainta cu miliarde de ani în viitor în capitolul 6, unde, paradoxal, putem trage concluzii mai clare decât în capitolele precedente, căci limitele ultime ale vieţii în cosmosul nostru nu sunt impuse de inteligenţă, ci de legile fizicii.

După ce vom fi încheiat explorarea istoriei inteligenţei, ne vom confrunta cu întrebările ce viitor este de dorit şi cum putem ajunge la el. Pentru a lega datele seci de problema scopului şi a semnificaţiei, vom examina bazele fizice ale obiectivelor în capitolul 7, iar conştiinţa în capitolul 8. în fine, în epilog, vom vedea ce poate fi făcut chiar acum pentru a contribui la edificarea viitorului pe care ni-l dorim.

Dacă îţi place să citeşti pe sărite, majoritatea capitolelor sunt relativ autonome odată ce ai deprins terminologia şi definiţiile din acest prim capitol şi de la începutul următorului. Dacă eşti



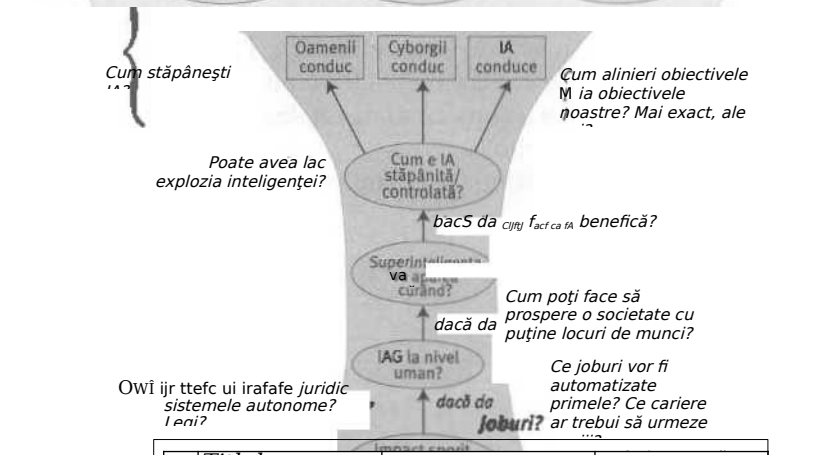


Figura 1.6 Înt
se ajunge la

	Titlul capitolului	Subiect	Caracter
	Preludiu: Povestea Echipei Omega	**Material pentru reflecţie**	**Extrem speculativ**
1.	**Dezbaterea**	**Idei-cheie, terminologie**	
2.	**Materia devine inteligentă**	**Bazele inteligenţei**	**Nu Foarte speculativ**
3.	**IA, economie, arme şi legi**	**Viitorul apropiat**	
4.	**Explozia inteligenţei?**	**Scenarii ale supe (inteligenţei**	
5-Ó.	**Urmări Zestrea noastră cosmică: Obiective**	**Următorii 10 000 de Următoarele miliarde de ani**	**Extrem speculativ**
7.	**Istoria**	**lui orientat spre**	**Nu Foarte speculativ**
8.	**Conţiinţa**	**Conţiinţa naturală şi artificială**	**Speculativ**
	Epilog: Povestea Echipei FU	**Ce artrebui să Facem?**	**Nu Foarte SPKIAtiV**

Figura 1.7 Structura cărţii

cercetător IA, poţi sări întreg capitolul *2*, cu excepţia definiţiilor de Ia început. Dacă eşti novice, capitolele 2 şi 3 îţi demonstrează de ce capitolele de la 4 la 6 nu pot fi simplu etichetate drept „poveşti SF imposibile". Figura 1.7 arată cât de speculativ este conţinutul diverselor capitole.

Ne aşteaptă o călătorie fascinantă. Să începem!

REZUMAT:

- Viaţa, definită ca proces careîşi păstrează ţi replichează complexitatea, poate cunoaşte trei etape: etapa biologică (i.o), în care hardware-ul şi software-ul ei evoluează, etapa culturală (2.0), în care îţi poate proiecta software-ul prin învăţare, şi etapa tehnologică (3.0), în care îşi poate proiecta şi hardware-ul, devenind stăpână pe destinul ei.
- Inteligenţa artificială ne poate permite să lansăm Viaţa 3.0m acest secol, şi s-a declanşat o dezbatere fascinantă privind viitorul pe care ni-l dorim şi calea de a ajunge la el. Trei tabere principale au intrat în dezbatere: tehno-sceptici, utopişti digitali şi mişcarea IA benefică.
- Tehno-scepticii cred că IAG supraumană e atât de dificil de creat, încât nu va apărea în următoarele sute de ani, deci e ridicol să ne preocupăm acum de ea (şi de Viaţa 3.0).
- Utopiştii digitali cred că e probabilă apariţia eiîn acest secol şi întâmpină cu inima deschisă Viaţa 3.0, considerând-o următorul pas, natural şi de dorit, în evoluţia cosmică.
- Mişcarea IA benefică crede de asemenea că e probabilă apariţia eiîn acest secol, dar nu consideră că rezultatul bun e garantat, ci că acesta trebuie asigurat printr-o muncă serioasă sub forma cercetării în siguranţa IA.
- în afară de aceste controverse legitime, în care experţii nu cad de acord, există şi pseudocontroverse plictisitoare provocate de confuzii. De pildă, nu-ţi pierde vremea discutând despre „viaţă“, „inteligenţă“ sau „conştiinţă“ până nu te asiguri că pentru tine şi interlocutorul tău aceste cuvinte înseamnă acelaşi lucru! Cartea de faţă foloseşte definiţiile din tabelul 1.1.
- Fereşte-te şi de afirmaţiile false din figura 1.5: „în 2100 superinteligenţa va fi inevitabilă/imposibilă“, „Numai ludiţîi sunt îngrijoraţi de IA”, „Ne îngrijorează faptul că IAardeveni rea şi/sau conştientă, şi că ne despart de ea doar câţiva ani”, „Roboţii sunt principala problemă“, „IA nu poate controla oamenii şi nu poate avea obiective“.
- în capitolele de la 2 la *6* vom explora povestea inteligenţei de la umilele ei începuturi cu miliarde de ani în urmă până la posibilele viitoruri cosmice de peste miliarde de ani. Vom analiza întâi provocările pe termen scurt, legate de joburi, arme IA şi încercarea de a ajunge la IAG la nivel uman, iar apoi un fascinant spectru de viitoruri posibile cu maşini inteligente şi/sau oameni. Mă întreb ce variantă preferaţi!
- în capitolele de la 7 la 9 vom trece de la descrierea rece factuală la explorarea obiectivelor, conştiinţei şi sensului, şi vom investiga ce

putem face chiar acum pentru a contribui la edificarea viitorului dorit.

- Consider că aceasta discuţie despre viitorul vieţii cu IA este cea mai importantă dezbatere a epocii noastre – aşa încât vă invit să participaţi!

CAPITOLUL 2

Materia devine inteligentă

Hidrogenulf...], dacă are suficient timp la dispoziţie, se transformă în oameni. — Edward Rabert Harrison, 1995

Una dintre cele mai spectaculoase evoluţii în cursul celor 13,8 miliarde de ani scurşi de la big bang este că acea materie mută şi neînsufleţită a devenit inteligentă. Cum s-a petrecut asta şi cum poate progresa inteligenţa în viitor? Ce are de spus ştiinţa despre istoria şi destinul inteligenţei în cosmosul nostru? Pentru a aborda aceste întrebări, vom examina în capitolul de faţă bazele inteligenţei şi elementele ei constitutive. Ce înseamnă că un bulgăre de materie e inteligent? Ce înseamnă că un obiect poate memora, calcula şi învăţa?

CE ESTE INTELIGENŢA?

Soţia mea şi cu mine am avut ocazia să asistăm la un simpozion despre inteligenţa artificială organizat de Fundaţia Suedeză Nobel, iar atunci când mai multor cercetători IA de vârf li s-a cerut să definească inteligenţa, s-au lansat într-o dispută lungă, fără să ajungă la consens. Ni s-a părut amuzant: nici măcar specialiştii în inteligenţa artificială nu cad de acord! Nu există aşadar

nici o definiţie „corectă“ indiscutabilă a inteligenţei. Există însă multe definiţii concurente, vizând, între altele, aptitudinea pentru logică, înţelegere, planificare, cunoaştere afectivă, conştiinţă de sine, creativitate, rezolvare a problemelor şi învăţare.

În explorarea viitorului inteligenţei, vrem să avem o perspectivă cât mai largă, fără să ne limităm la acele genuri de inteligenţă care există deocamdată. De aceea definiţia pe care am dat-o în primul capitol şi modul în care voi folosi cuvântul în întreaga carte vizează un spectru atât de amplu:

inteligenţa = *capacitatea de a îndeplini obiective complexe*

Definiţia e suficient de largă pentru a le include pe toate cele sus-menţionate, odată ce înţelegerea, conştiinţa de sine, rezolvarea problemelor, învăţarea etc. sunt toate exemple de obiective complexe. Este de asemenea suficient de largă pentru a subsuma definiţia din *Dicţionarul Oxford* – „capacitatea dea dobândi şi folosi cunoştinţe şi deprinderi“ –, fiindcă poţi avea drept obiectiv să foloseşti cunoştinţe şi deprinderi.

Pentru că există multe obiective posibile, există şi multe tipuri de inteligenţă posibile. Conform definiţiei noastre, n-are sens să cuantifici inteligenţa oamenilor, a animalelor sau a maşinilor printr-un singur număr cum e IQ.[††] Care e mai inteligent, un program de calculator care joacă numai şah, sau unul care joacă numai go? Nu există un răspuns rezonabil, căci sunt două lucruri care nu pot fi

†† Pentru a înţelege asta, închipuiţi-vă cum aţi reacţiona dacă cineva ar vrea să cuantifice cu un singur număr, „coeficientul atletic“, capacitatea de a se descurca la probele olimpice, aşa încât sportivul cu cel mai mare coeficient să câştige medaliile de aur la toate probele. *(N. a.)*

* Jeopardy! – joc dintr-o emisiune TV din America. (W. t)

comparate. Putem totuşi spune că mai inteligent e un al treilea program care se dovedeşte a fi cel puţin la fel de bun ca ele la îndeplinirea *tuturor* obiectivelor, şi strict mai bun în cel puţin o privinţă (să câştige la şah, de pildă).

Nu prea are sens nici să despici firul în patru pentru a stabili dacă ceva este sau nu inteligent în cazurile-limită, odată ce există un spectru al capacităţilor, nu e vorba aici de totul-sau-nimic. Ce oameni au capacitatea de a îndeplini obiectivul de a vorbi? Nou-născuţii?

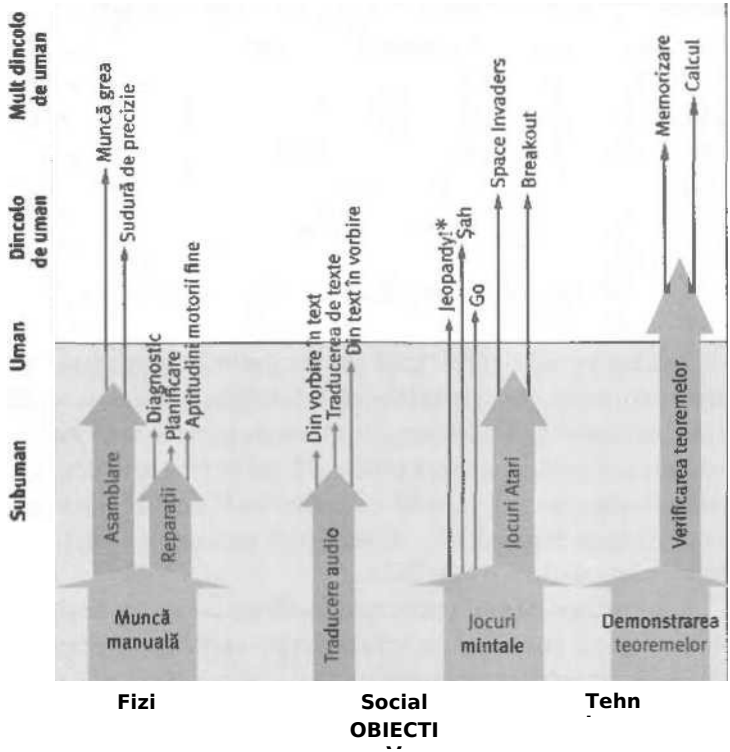


Figura 2.1 Inteligenţa, definită drept capacitatea de a îndeplini obiective complexe, nu poate fi măsurată printr-un singur IQ, ci doar printr-un spectru de capacităţi în funcţie de obiective. Fiecare săgeată arată cât de apte sunt azi cele mai bune sisteme IA pentru a îndeplini diferite obiective, indicând faptul că IA de azi este *îngustă,* fiecare sistem fiind capabil să îndeplinească doaranumite obiective. Dimpotrivă, inteligenţa umană are un spectru remarcabil de larg: un copil sănătos poate învăţa să fie superior în aproape toate privinţele.

Nu. Realizatorii de emisiuni radio? Da. Cum rămâne însă cu ţâncii care cunosc zece cuvinte? Sau cinci sute de cuvinte? Unde tragi linia de demarcaţie? Am folosit deliberat cuvântul „complex“ în definiţia de mai sus, fiindcă n-are rost să faci o delimitare artificială
între inteligenţă şi non-inteligenţă, ci e mai util să cuantifici gradele capacităţii de a îndeplini diferite

obiective.

Pentru a clasifica diversele inteligenţe într-o taxonomie, o altă distincţie importantă e cea între inteligenţa *îngustă* şi cea *largă.* Calculatorul IBM pentru şah Deep Blue, care l-a învins pe campionul mondial Gări Kasparov în 1997, nu putea îndeplini decât foarte îngusta sarcină de a juca şah – în ciuda unui hardware şi software impresionante, nu-l putea bate nici măcar pe un copil de patru ani la *„x* şi 0“. Sistemul IA DQN al lui DeepMind produs de Google poate îndeplini o gamă ceva mai largă de obiective: poate juca zeci de jocuri pe calculator Atari la nivel uman sau mai bine. Inteligenţa umană e însă incomparabil mai largă, având o mulţime de aptitudini. Un copil sănătos, având la dispoziţie suficient timp să se antreneze, se poate descurca destul de bine nu doar în *orice* jocuri, dar şi în orice limbă, sport sau meserie. Comparând inteligenţa de azi a oamenilor şi a maşinilor, oamenii câştigă detaşat la capitolul lărgimea spectrului, în timp ce maşinile ne depăşesc într-un mic număr, dar în creştere, de domenii înguste, după cum se vede în figura 2.1. Marea ţintă a cercetătorilor IA e să construiască JA generală“ (IAG), capabilă să îndeplinească practic orice obiectiv, inclusiv învăţarea. Vom vorbi despre asta în detaliu în capitolul 4. Termenul „IAG“ a fost popularizat de cercetătorii Shane Legg, Mark Gubrud şi Ben Goertzel cu sensul mai precis de inteligenţă artificială generală la *nivel umarr.* capacitatea de a îndeplini orice obiectiv cel puţin Ia fel de bine ca oamenii.[1] Voi folosi definiţia lor, aşadar, dacă nu califc explicit acronimul (scriind „IAG supraumană“, de pildă), prin „IAG“ am în vedere „IAG Ia nivel uman“.[##]

Unii preferă „LA la nivel uman“ sau „IA tare“ ca sinonime pentru IAG, dar ambele sunt problematice. Chiar şi un calculator de buzunar e la nivel uman intr-un sens îngust. Antonimul lui „IA tare“ este „IA slabă“, dar e ciudat să numeşti „slabe“ sisteme IA precum Deep Blue, Watson şi AlphaGo. *(N. a.)*

Deşi cuvântul „inteligenţă“ are de regulă o conotaţie pozitivă, trebuie observat că îl folosesc într-un sens complet neutru: capacitatea de a îndeplini obiective complexe, indiferent dacă ele sunt
considerate bune sau rele. 0 persoană inteligentă poate excela deci în a-i ajuta pe oameni sau în a-i răni. Mai este de făcut o distincţie subtilă în privinţa obiectivelor: cui aparţin? Să presupunem că viitorul tău secretar personal robot nu are obiective proprii, dar va face tot ce-i ceri, iar tu îi ceri să pregătească o cină italiană perfectă. Dacă intră pe internet şi caută reţete de mâncăruri italieneşti, cum să ajungă la cel mai apropiat supermarket, cum să strecoare pastele etc., iar apoi cumpără ingredientele şi prepară o cină gustoasă, îl vei considera pesemne inteligent, deşi scopul iniţial îţi aparţine. De fapt, el a adoptat obiectivul tău odată ce ai făcut cererea, pentru a-l diviza într-o ierarhie de sub- obiective, de la plata la magazin până la raderea parmezanului, în acest sens, comportamentul inteligent este inexorabil legat de atingerea obiectivului.

E firesc să evaluăm dificultatea sarcinilor după posibilităţile oamenilor, ca în figura 2.1, ceea ce nu mai e valabil pentru calculatoare. Pare să fie mult mai greu să înmulţeşti 314159 cu 271828 decât să recunoşti un prieten într-o fotografie, dar calculatoarele ne depăşeau clar la socoteli cu mult înainte să mă nasc, în vreme ce recunoaşterea imaginilor la nivel uman a devenit doar de curând posibilă. Sarcini senzomotorii de bază ni se par uşoare, deşi cer resurse computaţionale enorme – acesta e paradoxul lui Moravec, care se explică prin faptul că în creierul nostru pentru asemenea sarcini sunt alocate mari cantităţi de hardware specializat, peste un sfert din creier.

Mi-am luat libertatea să ilustrez în figura 2.2 următoarea metaforă a lui Hans Moravec

Calculatoarele sunt maşini universale, potenţialul lor se extinde uniform asupra unei întinderi nesfârşite de sarcini. Pe de altă parte, potenţialul uman e mare în domenii importante pentru supravieţuire, dar mic în domenii îndepărtate de aceasta, închipuiţi-vă un „peisaj al competenţei umane“, cu şesuri pentru „socoteli“ şi „memorare papagalicească“, dealuri pentru „demonstrarea teoremelor“ şi „jocul de şah“ şi munţi înalţi pentru „locomoţie“, „coordonare mână-ochi“ şi „interacţie socială“.

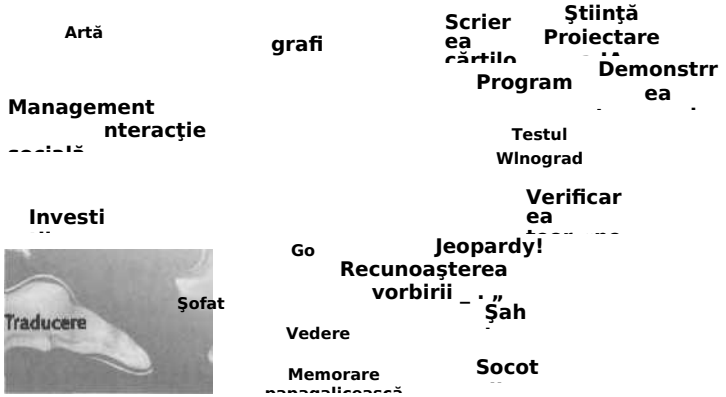


Figura 2.2 Ilustrarea „peisajului competenţei umane“ al lui Hans Moravec.în care altitudinea reprezintă dificultatea pentru calculatoare, iar nivelul în creştere al apei reprezintă ce pot face calculatoarele.

> Progresul performanţelor calculatoarelor este ca apa care inundă încet peisajul. Cu o jumătate de secol în urmă a înecat şesurile, înlocuind socotitorii şi arhivarii, dar lăsându-ne us- caţi pe cei mai mulţi dintre noi. Acum apa a ajuns la poalele dealurilor, iar avanposturile noastre bat în retragere. Ne simţim siguri pe crestele munţilor, dar în ritmul actual şi ele vor fi sub ape peste încă o jumătate de secol. Propun să ne construim arce când acea zi se va apropia, şi să adoptăm o viaţă de navigatori?

în deceniile de când Moravec a scris aceste rânduri, nivelul mării a crescut continuu, după cum a prezis, unele dealuri (inclusiv şahul) au fost inundate, iar apa ar putea ajunge într o bună zi la un punct critic, declanşând schimbări spectaculoase. Nivelul critic este cel în care maşinile încep să poată proiecta IA. Până atunci, creşterea nivelului mării e provocată de *oameni* care perfecţionează maşini; după aceea, creşterea poate fi controlată de *maşini* care perfecţionează maşini, fiind pesemne mult accelerată şi inundând rapid întregul uscat. Aceasta e fascinanta şi controversata idee de

singularitate, despre care vom vorbi în capitolul 4.

Alan Turing, unul dintre pionierii informaticii, a demonstrat că dacă un calculator poate efectua un anumit set minim de operaţii, atunci, având la dispoziţie timp şi memorie suficiente, poate fi programat să facă tot ceea ce *oricare* alt calculator poate face. Maşinile care depăşesc acest prag critic se numesc calculatoare universale; toate telefoanele inteligente şi laptopurile de azi sunt universale în acest sens. în mod analog, cred că pragul critic de inteligenţă necesară pentru a proiecta IA este pragul *inteligenţei universale.* având la dispoziţie timp şi resurse suficiente, se poate face pe sine capabilă să îndeplinească orice obiectiv la fel de bine ca *oricare* altă entitate inteligentă. De pildă, dacă se hotărăşte să-şi îmbunătăţească aptitudinile sociale, de previziune sau de proiectare IA, o poate face. Dacă se hotărăşte să descopere cum se construieşte o fabrică de roboţi, o poate face. Cu alte cuvinte, inteligenţa universală poate deveni Viaţa 3.0.

în ultimă instanţă, inteligenţa ţine de informaţie şi calcul, nu de came, sânge sau atomi de carbon – ceea ce înseamnă că nu există motive fundamentale pentru care maşinile să nu devină într-o bună zi cel puţin la fel de inteligente ca noi. Dar ce sunt de fapt informaţia şi calculul, odată ce fizica ne spune că la nivel fundamental totul e doar materie şi energie? Cum poate ceva atât de abstract, impalpabil şi eteric precum informaţia şi calculul să fie încorporat în entităţi fizice palpabile? în particular, cum poate manifesta comportamentul pe care-l numim inteligent o grămadă de particule mişcându-se după legile fizicii?

Dacă răspunsul ţi se pare evident şi consideri plauzibil ca maşinile să devină la fel de inteligente ca oamenii în acest secol – fiindcă, de pildă, eşti cercetător IA –, poţi sări direct la capitolul 3. în caz contrar, află că am scris

următoarele trei secţiuni special pentru tine.

CE ESTE MEMORIA?

Când spunem că un atlas conţine *informaţie* despre lume, înţelegem ca există o relaţie între starea cărţii (în particular, poziţiile anumitor molecule care dau culoare literelor şi imaginilor) şi starea lumii (de pildă, poziţiile continentelor). Dacă continentele s-ar afla în locuri diferite, atunci şi acele molecule s-ar afla în locuri diferite. Noi, oamenii, folosim o gamă largă de dispozitive pentru stocarea informaţiei, de Ia cărţi şi creiere la discuri dure, toate având această proprietate: starea lor poate fi pusă în corespondenţă cu (deci ne poate informa despre) starea altor lucruri care ne interesează.

Ce proprietate fizică fundamentală, care le face utile ca dispozitive de memorie (de stocare a informaţiei) au ele în comun? Răspunsul este că se pot afla în stări diferite de durată lungă – suficient de lungă pentru a codifica informaţia până ai nevoie de ea. Ca exemplu simplu, să presupunem că aşezi o bilă pe o suprafaţă deluroasă cu 16 văi, ca în figura 2.3. După ce bila se rostogoleşte şi se opreşte, ea se va afla într-unul dintre cele 16 locuri, aşa încât poţi folosi poziţia ei pentru a ţine minte orice număr de la 1 la 16.

Acest dispozitiv de memorie e destul de rezistent, fiindcă chiar dacă e puţin scuturat de forţe externe, bila e probabil să rămână în aceeaşi vale în care ai aşezat-o, deci ştii mai departe ce număr e stocat. Motivul pentru care această memorie e atât de stabilă este că ridicarea bilei din vale necesită mai multă energie decât e probabil să ofere perturbaţiile aleatorii. Aceeaşi idee poate sta la baza unor memorii mult mai generale decât în cazul unei bile care se rostogoleşte: energia unui sistem fizic complicat poate depinde de tot felul de proprietăţi fizice, chimice, electrice şi magnetice, iar câtă vreme e nevoie de energie pentru a scoate sistemul din starea pe care

vrei s-o memorezi, acea stare va fi stabilă. De aceea solidele au multe stări cu durată lungă de viaţă, dar nu şi lichidele şi gazele: dacă gravezi numele cuiva pe un inel de aur, informaţia va rămâne acolo ani de zile, fiindcă remodelarea aurului presupune multă energie, dar dacă scrii numele pe suprafaţa unui lac, se pierde instantaneu, căci suprafaţa apei îşi schimbă fără efort forma.

Cel mai simplu dispozitiv de memorie are doar două stări stabile (figura 2.3). Ne putem deci închipui că el codifică un număr binar (un „bit"), adică un zero sau un unu. Informaţia stocată de orice dispozitiv mai complicat poate fi în mod echivalent stocată în biţi multipli: de pildă, împreună, cei patru biţi din figura 2.3 (dreapta) se pot afla în 2 x 2 x 2 x *2* = 16 stări diferite, adică 0000, 0001,0010... 1111, deci au exact aceeaşi capacitate de memorare ca mai complicatul sistem cu 16 stări (din stânga). Putem privi aşadar biţii ca pe atomi de informaţie – bucata cea mai mică de informaţie care nu mai poate fi divizată şi care se poate combina pentru a alcătui orice informaţie. De pildă, tocmai am tastat cuvântul „word", iar laptopul meu l-a memorat ca secvenţa de 4 numere 119111114100, stocând fiecare număr ca 8 biţi (fiecare literă mică e reprezentată printr-un număr egal cu 96 plus ordinea ei în alfabet). Când ating tasta *w,* laptopul îmi afişează pe ecran litera u>, iar această imagine e de asemenea reprezentată prin biţi: 32 de biţi specifică culoarea fiecăruia dintre milioanele de pixeli ai ecranului.

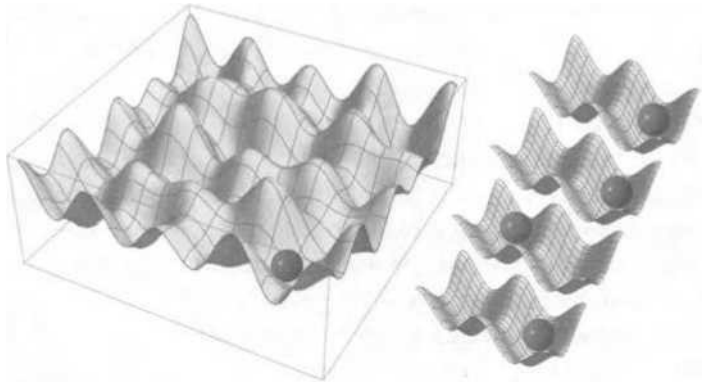


Figura 2.3 Un sistem fizic e un dispozitiv de memorie util dacă se poate afla în mai multe stări stabile. Bila din stânga poate codifica patru biţi de informaţie desemând în care dintre cele 2^* «16 văi se află. împreună, cele patru bile din dreapta codifică şi ele patru biţi de informaţie – fiecare câte un bit.

Din moment ce sistemele cu două stări sunt uşor de produs şi de manevrat, majoritatea calculatoarelor modeme stochează informaţia ca biţi, însă aceşti biţi sunt realizaţi intr-o gamă largă de moduri. Pe un DVD, fiecare bit corespunde prezenţei sau absenţei unei adâncituri microscopice într-un anumit loc pe suprafaţa de plastic. Pe un disc dur, fiecare bit corespunde unui punct de pe suprafaţă care e magnetizat într-unul din două moduri. în memoria de lucru a laptopului meu, fiecare bit corespunde poziţiei anumitor electroni, determinând dacă un dispozitiv numit microcondensator e încărcat. Unele tipuri de biţi pot fi transportate, chiar şi cu viteza luminii, ca într-o fibră optică ce transmite e-mailul tău, fiecare bit corespunzând unui fascicul laser mai intens sau mai slab la un moment dat.

Inginerii preferă să codifice biţii în sisteme care sunt nu doar stabile şi de pe care e uşor de citit (ca un inel de aur), ci pe care e şi uşor de scris: modificarea stării discului tău dur necesită mult mai puţină energie decât gravarea aurului. Ei preferă de asemenea sistemele cu care e comod de lucrat şi care sunt ieftine. Pe de altă

parte însă, nu le pasă cum sunt reprezentaţi biţii ca obiecte fizice – după cum nu-i pasă nici utilizatorului, fiindcă pur şi simplu nu contează! Dacă-i trimiţi prin e-mail prietenului tău un document să-l imprime, informaţia poate fi copiată într-o succesiune rapidă din magnetizările de pe discul tău dur în sarcinile electrice ale memoriei de lucru a calculatorului, undele radio din reţeaua wireless, tensiunile din router, pulsurile laser dintr-o fibră optică, şi în cele din urmă în moleculele unei foi de hârtie. Cu alte cuvinte, *informaţia are propria ei viaţă, independentă de substratul eifizid* De regulă, ne interesează doar acest aspect al informaţiei independent de substrat: dacă prietenul te sună să discutaţi despre documentul pe care i l-ai trimis, pesemne că nu vrea să vorbească despre tensiuni sau molecule. Acesta e primul nostru indiciu privind felul în care ceva impalpabil precum inteligenţa poate fi încorporat în ceva palpabil, şi vom vedea curând că ideea independenţei de substrat e mult mai profundă, incluzând nu doar informaţia, ci şi calculul şi învăţarea.

Datorită acestei independenţe de substrat, inginerii au putut înlocui în repetate rânduri dispozitivele de memorie din calculatoare cu unele tot mai bune, bazate pe noi tehnologii, fără să fie nevoie de schimbări în software. Rezultatul a fost spectaculos, după cum se vede în figura *2.4; în* ultimele şase decenii, preţul memoriei calculatoarelor s-a înjumătăţit cam la fiecare doi ani. Discurile dure au devenit de peste 100 de milioane de ori mai ieftine, iar memoriile rapide, utile în calcule, nu doar pentru stocare, au devenit de 10 bilioane de ori mai ieftine. Dacă ai obţine o asemenea reducere de 99,99999999999% la toate cumpărăturile, ai putea cumpăra toate imobilele din New York cu vreo 10 centi, şi tot aurul care a fost vreodată extras cu un dolar.

Pentru mulţi dintre noi, aceste îmbunătăţiri spectaculoase ale tehnologiei memoriei se leagă de poveşti personale. Ţin minte că

TerabytelO[1’]
10”
10[w]
Gigabyte 10?
10*
10’
Kilobyte 10[6]
10?
IC
10?
10?
10
Byte 1
0.1
0.01
0.001

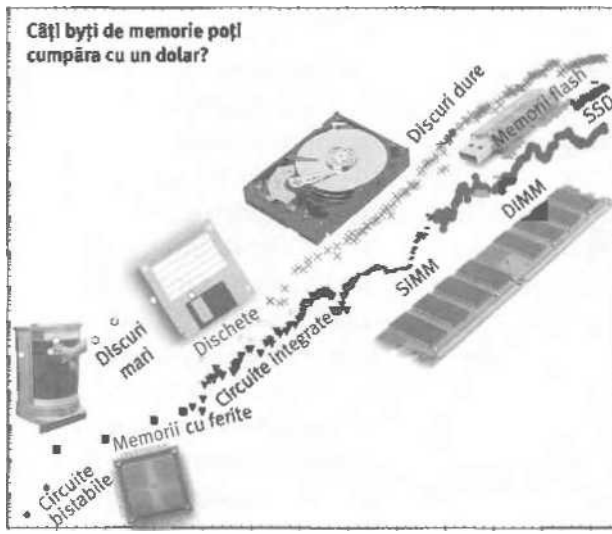


Figura *2.4* în ultimele şase decenii, preţul memoriei calculatoarelor s-a înjumătăţit cam (a fiecare doi ani, corespunzând unui factor de ieftinire de o mie cam la fiecare douăzeci de ani. Un byte reprezintă opt biţi. Date obţinute prin amabilitatea lui John McCallum din http//www.jcmit.net/memoryprice.htm.

în timpul liceului am muncit într-un magazin de dulciuri ca să-mi cumpăr un calculator cu o memorie de 16 kilobyţi, iar atunci când am făcut şi am vândut un procesor de text pentru el împreună cu colegul meu de clasă Magnus Bodin, a trebuit să-l scriem într-un cod-maşină ultracompact, ca să lăsăm suficientă memorie pentru textul care trebuia procesat. După ce m-am obişnuit cu dischetele de 70 kB, am fost uimit de dischetele mai mici, de 3,5 ţoii, care stocau 1,44 MB, putând conţine o carte întreagă, pentru a descoperi apoi discurile dure de 10 MB – pe care abia încape un cântec din descărcările de azi. Aceste amintiri din adolescenţă mi s-au părut aproape ireale ieri, când am cumpărat cu vreo sută de dolari un disc dur cu o capacitate de 300 000 de ori mai mare.

Ce putem spune despre dispozitivele de memorie care n-au fost proiectate de oameni, ci au evoluat? Biologii nu

ştiu deocamdată care a fost prima formă de viaţă care şi-a copiat structura de-a lungul generaţiilor, dar trebuie să fi fost foarte mică. O echipă condusă de Philipp Holliger de la Universitatea Cambridge a creat în 2016 o moleculă de ARN care coda 412 biţi de informaţie genetică şi putea copia secvenţe de ARN mai lungi decât ea însăşi – ceea ce venea în sprijinul ipotezei „lumii ARN", după care viaţa primitivă pe Pământ a implicat scurte fragmente de ARN autoreplicatoare. Cel mai mic dispozitiv de memorie cunoscut care a evoluat este genomul bacteriei *Candidatus Carsonella ruddii,* stocând circa 40 de kilobyţi, în vreme ce ADN-ul uman stochează circa 1,6 gigabyţi, comparabil cu un film descărcat. După cum am spus în capitolul precedent, creierul nostru stochează mult mai multă informaţie decât genele noastre; electric aproximativ 10 gigabyţi (specificând care dintre cei 100 de miliarde de neuroni sunt activaţi în orice moment) şi chimic/biologic 100 de terabyţi (specificând cât de puternic sunt legaţi diverşi neuroni prin sinapse). Compararea acestor cifre cu cele ale memoriilor maşinilor arată că cele mai bune calculatoare din prezent pot depăşi orice sistem biologic – Ia un preţ care în 2016 era de câteva mii de dolari, şi care scade rapid.

Memoria din creier funcţionează diferit de cea din calculator, atât prin felul în care e construită, cât şi prin felul în care e folosită. Dacă accesezi datele dintr-un calculator specificând *unde* sunt stocate, regăseşti amintirile din creier specificând *ce anume* e stocat. Fiecare grup de biţi din memoria calculatorului are o adresă numerică, iar pentru a găsi un fragment de informaţie, calculatorul specifică la ce adresă să cauţi, la fel cum ţi-aş spune eu „Du-te la biblioteca mea, ia a cincea carte din dreapta de pe raftul de sus şi spune-mi ce scrie Ia pagina 314". în schimb, regăseşti informaţia din creier la fel cum foloseşti un motor de căutare: specifici un fragment de informaţie sau ceva legat de ea, şi informaţia apare. Dacă îţi spun „a fi sau a nu", sau dacă tastez asta

pe Google, sunt şanse să ajung la „A fi sau a nu fi, aceasta-i întrebarea“. Metoda funcţionează probabil şi dacă folosesc altă parte din citat sau amestec lucrurile. Asemenea sisteme de memorie se numesc *auto-asociative,* fiindcă ne amintim prin asociaţie, nu prin adresă.

într-un celebru articol din 1982, fizicianul John Hopfield arăta că o reţea de neuroni interconectaţi poate funcţiona ca o memorie auto-asociativă. Ideea de bază mi se pare foarte frumoasă, şi e valabilă pentru orice sistem fizic cu stări stabile multiple. De pildă, să considerăm o bilă pe o suprafaţă cu două văi, ca sistemul cu un bit din figura 2.3, şi să modelăm suprafaţa aşa încât coordonatele *x* ale celor două puncte de minim unde bila se află în repaus sunt x» 72*1,41421 şi respectiv n ~ 3,14159. Dacă ţii minte doar că TI este aproape de 3, pui pur şi simplu bila la x = 3 şi vezi cum se apropie de o valoare mai exactă a lui n rostogolindu-se spre minimumul cel mai apropiat. Hopfield a înţeles că o reţea complexă de neuroni oferă un peisaj asemănător, cu foarte multe valori minime ale energiei în care sistemul se poate stabiliza, şi s-a demonstrat mai târziu că poţi comprima 138 de amintiri diferite la fiecare o mie de neuroni, fără a provoca mari confuzii.

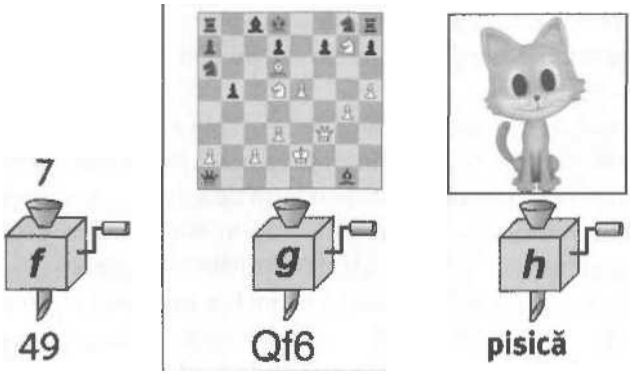
CE ESTE CALCULUL?

Am văzut ca un obiect fizic poate memora informaţie. Dar cum poate el calcula?

Un calcul e o transformare a stării unei memorii în altă stare.

Cu alte cuvinte, calculul ia informaţia şi o transformă, punând

în aplicare ceea ce matematicienii numesc *o funcţie.* Privesc funcţia ca pe un fel de maşina de tocat came pentru informaţie, după cum e ilustrat în figura 2.5: pui informaţia sus, roteşti de manivelă şi obţii informaţia procesată jos – şi poţi repeta asta de câte ori vrei, cu diferite input-uri. Procesarea informaţiei e determi- nistă, în sensul că, dacă o repeţi cu acelaşi input, obţii mereu acelaşi output.



Agura 2.5 Un *calcul* ia informaţia şi o transformă, punând în aplicare ceea ce matematicienii numesc *funcţie.* Funcţia *f* (din stânga) ia biţi reprezentând un număr şi îi ridică la pătrat. Funcţia *g* (din mijloc) ia biţi reprezentând o poziţie de şah şi calculează cea mai bună mutare pentru alb. Funcţia *h* (din dreapta) ia biţi reprezentând o imagine şi calculează un nume care o descrie.

Deşi descriptiv sună simplu, ideea de funcţie e incredibil de generală. Unele funcţii sunt banale, cum este cea numită *NOT,* la care input-ul e un singur bit, iar output-ul inversul lui, transformân- du-l astfel pe zero în unu şi viceversa. Funcţiile pe care le învăţăm la şcoală corespund de regulă butoanelor unui calculator de bu- zunar, introducând unul sau mai multe numere şi obţinând un singur număr – de pildă, funcţia x^z îţi dă numărul introdus înmulţit cu el însuşi. Alte funcţii pot fi extrem de complicate. De exemplu, dacă dispui de o funcţie la care input-ul sunt biţii reprezentând o poziţie de

şah arbitrară, iar output-ul sunt biţii reprezentând cea mai bună mutare următoare posibilă, poţi să câştigi Campionatul Mondial de Şah pe Calculator. Daca ai o funcţie la care input-ul sunt toate datele financiare dii. lume, iar output-ul sunt cele mai bune acţiuni de cumpărat, vei deveni repede foarte bogat. Mulţi cercetători IA se ocupă în întreaga lor carieră cu punerea în aplicare a anumitor funcţii. De exemplu, obiectivul cercetărilor de traducere automată este de a pune în aplicare o funcţie ai căror biţi de input reprezintă un text într-o limbă, iar output-ul, un text în altă limbă, în vreme ce obiectivul cercetărilor de etichetare automată este ca pentru biţi de input prezentând o imagine să se obţină biţi de output reprezentând un text care o descrie (figura 2.5, dreapta).

Cu alte cuvinte, dacă poţi pune în aplicare funcţii foarte complexe, poţi construi o maşină inteligentă capabilă să îndeplinească obiective foarte complexe. Se reformulează astfel în termeni mai precişi întrebarea de la începutul capitolului: cum poate o bucată de materie neînsufleţită calcula o funcţie complicată?

în loc să rămână imobilă ca un inel de aur sau alt dispozitiv de memorie static, ea trebuie să aibă o *dinamică* complexă, aşa încât starea ei viitoare să depindă de starea actuală într-un mod complicat (şi, e de dorit, controlabil/programabil). Aranjamentul ei atomic trebuie să fie mai puţin ordonat decât cel al unui solid rigid, în care nu se modifică nimic interesant, dar mai ordonat decât cel al unui lichid sau al unui gaz. Mai precis, vrem ca sistemul să aibă proprietatea ca, dacă îl punem într-o stare ce codifică informaţia de input, s-o lase un timp să evolueze conform legilor fizicii, apoi să interpreteze starea finală ca pe informaţia de output, iar out- put-ul să fie funcţia dorită a input-ului. Dacă aşa se întâmplă, atunci putem spune ca sistemul calculează funcţia noastră.

Ca prim exemplu, să examinăm cum putem construi o funcţie foarte simplă (dar importantă) numită *poartă NAND*[§§] din materie neînsufleţită. Input-ul e format din doi biţi, output-ul dintr-unul: output-ul e o dacă ambii biţi de input sunt 1; în toate celelalte cazuri, output-ul e i. Dacă conectăm două comutatoare în serie cu o baterie şi un electromagnet, prin electromagnet va circula curent doar dacă primul comutator şi al doilea sunt închise. Să plasăm un al treilea comutator sub electromagnet, ca în figura 2.6, aşa încât electromagnetul îl va deschide atunci când circulă curent prin el. Dacă interpretăm primele două comutatoare ca pe biţii de input, iar pe al treilea ca pe bitul de output (cu o = comutator deschis şi i = comutator închis), avem o poartă NAND: al treilea comutator e deschis doar dacă primele două sunt închise. Există multe alte moduri mai practice de a construi porţi NAND – de pildă, folosind tranzistori (figura 2.6). în calculatoarele de azi porţile NAND sunt de regulă constituite din tranzistori microscopici şi alte componente care pot fi automat gravate pe plăci de siliciu.

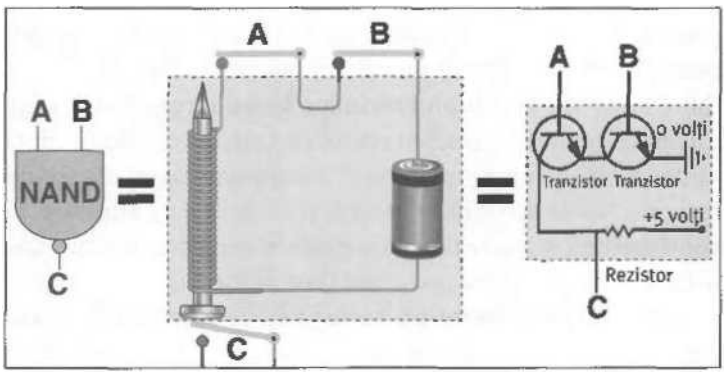


Figura 2.6 0 aşa-numită poartă NAND primeşte ca input doi biţi, A şi B, şi calculează ca output un bit, C, după regula că C – o dacă A-B-i,

§§ NAND e prescurtarea de la NOT AND („nu şi“). O poartă AND dă ca output 1 dacă primul input e 1 şi al doilea input e 1, deci NAND face exact pe dos. *(N. a.)*

şiC«i în celelalte cazuri. Numeroase sisteme fizice pot fi folosite ca porţi NAND. în exemplul din mijloc, comutatoarele sunt interpretate ca biţi (o – deschis, 1 = închis), iar când A şi B sunt închise, un electromagnet deschide comutatorul C.în exemplul din dreapta, tensiunile electrice sunt interpretate ca biţi (1 = cinci volţi, o = zero volţi), iar atunci când potenţialul în A şi B este cinci volţi, cei doi tranzistori conduc electricitatea, iar potenţialul în C scade aproximativ la zero volţi.

Există în informatică o remarcabilă teoremă care spune că porţile NAND sunt *universale,* în sensul că poţi pune în aplicare orice funcţie bine definită doar conectând porţi NAND.*** Dacă poţi construi aşadar suficiente porţi NAND, poţi crea un dispozitiv care să calculeze orice! Pentru a vă da o idee despre asta, în figura 2.7 arăt cum poţi înmulţi numere folosind numai porţi NAND.

Cercetătorii Norman Margolus şi Tommaso Toffoli de la MIT au creat termenul *computroniu* pentru a desemna orice substanţă care poate efectua calcule. Singura condiţie pentru computroniu este, după cum am văzut, să poată pune în funcţiune porţi NAND conectate oricum ne dorim, şi există o mulţime de tipuri de computroniu. 0 altă variantă simplă este de a înlocui porţile NAND cu porţi NOR, care dau ca output 1 doar dacă ambele input-uri sunt o. în secţiunea următoare vom vorbi despre reţelele neuronale, care pot de asemenea efectua calcule, deci acţionează ca un computroniu. Stephen Wolfram a arătat că acelaşi lucru e valabil pentru dispozitive simple numite automate celulare, care îşi modifică biţii în funcţie de biţii vecinilor. încă din 1936 Alan Turing a demonstrat intr-un celebru articol că o maşină simplă (numită acum „maşină

*** Prin „funcţie bine definită“ înţeleg ceea ce matematicienii şi infonna- ticienii numesc „funcţie calculabilă” – o funcţie care poate fi calculată cu un computer ipotetic, având memorie şi timp nelimitate. Alan Turing şi Alonzo Church au demonstrat că există funcţii care pot fi descrise, dar nu sunt cal- culabile. (N. o.)

Turing universală“) care poate manipula simboluri pe o bandă poate de asemenea efectua calcule arbitrare. în concluzie, nu numai că materia poate efectua orice calcul bine definit, dar o poate face într-o mulţime de moduri.

După cum am menţionat, Turing a demonstrat în acel articol din 1936 şi ceva mai profund: dacă un calculator poate efectua un anumit set de operaţii simple, atunci e *universal,* în sensul că, având suficiente resurse, poate face tot ce face orice alt calculator. A demonstrat că maşina Turing e universală, iar noi tocmai am văzut că această familie de calculatoare universale include obiecte diverse precum o reţea de porţi NAND şi o reţea de neuroni

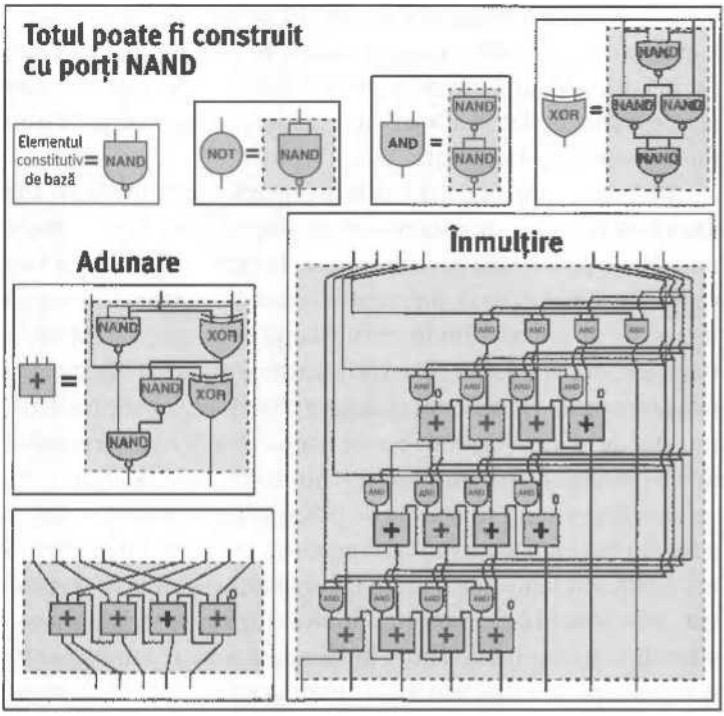


Figura 2,7 *Orice* calcul bine definit poate fi efectuat combinând ingenios doar porţi NAND. De pildă, modulele de adunare şi înmulţire de mai sus primesc ca input două numere binare reprezentate prin 4 biţi şi dau ca output un număr binar reprezentat prin 5 biţi şi, respectiv, 8 biţi. Modulele mai mici NOT, AND, XOR şi + (care adună trei biţi într-un număr binar de *2* biţi) sunt, la rândul lor, construite cu porţi NAND. înţelegerea deplină a acestei figuri e dificilă şi inutilă pentru lectura cărţii de faţă. Am inclus-o aici doar ca să ilustrez ideea de universalitate – şi din pedanterie.

interconectaţi. Stephen Wolfram a arătat că *majoritatea* sistemelor fizice nebanale, de la sisteme climatice la creiere, ar putea fi calculatoare universale dacă ar avea dimensiuni arbitrar de mari şi un timp de viaţă lung.

Faptul că exact acelaşi calcul poate fi efectuat pe *orice* calculator universal arată că într-adevăr *calculul e independent de substrat*- la fel ca informaţia, are propria

lui viaţă. Dacă ai fi un personaj conştient superinteligent dintr-un joc pe calculator din viitor, n-ai avea cum să-ţi dai seama dacă rulezi pe un desktop Windows, pe un laptop Mac OS sau pe un telefon Android, fiindcă ai fi independent de substrat. N-ai avea cum să ştii nici ce tip de tranzistori foloseşte microprocesorul.

Am înţeles importanţa ideii de independenţă de substrat când am văzut formele interesante sub care apare în fizică. De pildă, undele: au proprietăţi precum viteza, lungimea de undă şi frecvenţa – iar noi, fizicienii, putem studia ecuaţiile cărora li se supun fără să fie nevoie să ştim în *ce* substanţă se propagă acele unde. Când auzim ceva, detectăm undele acustice provocate de moleculele care se mişcă în amestecul de gaze numit aer, şi putem calcula tot felul de lucruri legate de aceste unde – cum scade intensitatea lor cu pătratul distanţei, cum îşi schimbă direcţia când trec prin uşi deschise şi cum se reflectă de pereţi generând ecoul – fără să ştim din ce e alcătuit aerul. De fapt, nici n avem nevoie să ştim că e alcătuit din molecule, putem ignora toate detaliile despre oxigen, azot, dioxid de carbon etc., fiindcă singura proprietate a substratului care contează şi intră în faimoasa ecuaţie a undelor este un număr pe care îl putem măsura: viteza undei, care, în acest caz, este de aproximativ 300 de metri pe secundă. Această ecuaţie a undelor, pe care am predat-o primăvara trecută studenţilor mei de la MIT, a fost descoperită şi folosită cu mult înainte ca fizicienii să ştie de existenţa atomilor şi moleculelor!

Acest exemplu ilustrează trei aspecte importante. în primul rând, independenţa de substrat nu înseamnă că substratul nu-i necesar, ci că majoritatea detaliilor lui nu contează. Evident, nu poţi avea unde acustice într-un gaz dacă nu există gazul, dar e suficient să existe un gaz oarecare. în mod asemănător, nu poţi avea calcul fără

materie, dar orice materie e bună, cu condiţia să poată fi aranjată în porţi NAND, să poată conecta neuroni sau alte elemente constitutive de bază care permit calculul universal. în al doilea rând, fenomenul independenţei de substrat îşi are propria lui viaţă. O undă poate străbate un lac chiar dacă nici una dintre moleculele de apă n-o face – ele doar urcă şi coboară, la fel ca „valurile“ din tribunele stadioanelor. în al treilea rând, adesea ne interesează numai aspectul independent de substrat: pe un surfer îl preocupă poziţia şi înălţimea valului, nu şi compoziţia lui moleculară. Am văzut că acest lucru e valabil în cazul informaţiei, iar el este valabil şi pentru calcul: dacă doi programatori încearcă să găsească eroarea din codul lor, probabil că nu discută despre tranzistori.

Am ajuns acum să răspundem Ia întrebarea cum se face că ceva fizic palpabil poate da naştere la ceva ce pare atât de impalpabil, abstract şi eteric precum inteligenţa: pare non-fizic pentru că e independent de substrat, are propria lui viaţă care nu depinde de detaliile fizice şi nu le reflectă. Pe scurt, calculul e un tipar în aranjamentul spaţio-temporar al particulelor, iar ceea ce contează e tiparul, nu particulele.

Cu alte cuvinte, hardware-ul e materia, iar software-ul e tiparul Această independenţă de substrat a calculului arată că IA e posibilă; inteligenţa n-are nevoie de carne, sânge sau atomi de carbon.

Datorită acestei independenţe de substrat, ingineri destoinici au putut înlocui în repetate rânduri tehnologiile din computerele noastre cu unele mult mai bune, fără a schimba software-ul. Rezultatul a fost la fel de spectaculos ca în cazul dispozitivelor de memorie. Aşa cum se vede în figura 2.8, preţul calculului s-a tot înjumătăţit cam la fiecare doi ani timp de peste un secol, ajungând să fie de un milion de milioane de milioane (10^{18}) de ori mai ieftin decât pe vremea când s-au născut

bunicii mei. Dacă totul s-ar ieftini de un milion de milioane de milioane de ori, atunci ai putea cumpăra cu o sutime de cent toate bunurile şi serviciile produse pe Pământ anul acesta. Scăderea spectaculoasă a costurilor este desigur motivul principal pentru care calculul a pătruns acum peste tot, trecând de la centrele de calcul de dimensiunea unei clădiri de mai ieri în casele, maşinile şi buzunarele noastre – ajungând până şi în locuri neaşteptate cum e încălţămintea sport.

De ce tehnologia noastră îşi dublează puterea la intervale regulate, prezentând ceea ce matematicienii numesc creştere exponenţială? De ce se întâmplă aşa nu doar cu miniaturizarea

Figura 2.8

începând din 1900, preţul calculului s-a tot înjumătăţit cam la fiecare doi ani. Acest grafic prezintă puterea de calcul măsurată în operaţii în virgulă mobilă pe secundă *(floating-point operations persecond* – FLOPS) care poate fi cumpărată cu 1000 de dolari.[3] Calculul care defineşte o operaţie în virgulă mobilă corespunde cu circa 105 operaţii logice elementare precum inversarea unui bit sau o evaluare NAND.

tranzistorilor (tendinţă cunoscută drept *legea lui Moare),* dar şi cu calculul în ansamblul lui (figura 2.8), cu memoria (figura 2.4) şi cu o mulţime de alte tehnologii, de la secvenţializareagenomului la imagistica cerebrală? Ray Kurzweil numeşte acest fenomen de dublare constantă „legea beneficiilor accelerate“.

Toate exemplele de dublare constantă pe care le cunosc din natură au aceeaşi cauză fundamentală, iar cel tehnologic nu face excepţie: fiecare pas îl creează pe următorul. De pildă, tu însuţi ai cunoscut o creştere exponenţială imediat ce ai fost conceput: fiecare celulă s-a divizat, dând naştere zilnic cam la două celule şi făcând ca numărul tău total de celule să crească zi de zi ca 1, *2,*4,

8,16 şi aşa mai departe. Conform celei mai plauzibile teorii ştiinţifice privind originea noastră cosmică, numită *inflaţie,* universul nostru a crescut Ia început Ia fel ca tine, dublându-şi în mod repetat dimensiunea la intervale regulate, până când un grăunte mult mai mic şi mai uşor decât un atom a devenit mai mare decât toate galaxiile pe care le-am văzut vreodată cu telescoapele noastre. Cauza a fost din nou un proces în care fiecare pas l-a declanşat pe următorul. La fel progresează şi tehnologia: odată ce tehnologia devine de două ori mai puternică, ea poate fi adesea folosită pentru a proiecta şi construi o tehnologie de două ori mai puternică, dublarea continuând în spiritul legii lui Moore.

Dar ceva care se dublează regulat, precum puterea tehnologiei noastre, ne-am putea gândi că trebuie să se oprească Ia un moment dat. Desigur, legea lui Moore nu se va mai aplica, fiindcă există o limită fizică în miniaturizarea tranzistorilor. Unii presupun în mod greşit că legea iui Moore e sinonimă cu dublarea con stantă a puterii noastre tehnologice. Ray Kurzwel arată însă că legea lui Moore implică nu prima, ci a cincea paradigmă tehnologică pentru a provoca creşterea exponenţială a calculului din figura 2.8: de fiecare dată când tehnologia a încetat să se perfecţioneze, ea a fost înlocuită de una mai bună. Când n-am mai putut micşora tuburile electronice, le-am înlocuit cu tranzistori, iar apoi cu circuite integrate, în care electronii se deplasează în două dimensiuni. Când această tehnologie îşi atinge limitele, există multe alte variante pe care le putem încerca – de pildă, să folosim circuite tridimensionale şi să punem la lucru altceva decât electronii.

Nimeni nu ştie cu certitudine care va fi următorul substrat computaţional, dar ştim că suntem încă departe de limitele impuse de legile fizicii. Seth Lloyd, colegul meu de la MIT, a calculat care e această limită

fundamentală şi, după cum vom vedea în detaliu în capitolul 6, ea se află cu 33 de ordine de mărime (io^{33}) dincolo de ceea ce poate face azi o bucată de materie care calculează. Aşa încât, chiar dacă dublăm mai departe puterea calculatoarelor la fiecare doi ani, vor mai trece două secole până să ajungem la frontiera ultimă.

Deşi toate calculatoarele universale pot efectua aceleaşi calcule, unele sunt mai eficiente decât altele. De pildă, un calcul care presupune milioane de înmulţiri nu necesită milioane de module separate de înmulţire construite din tranzistori diferiţi, ca în figura 2.6, ci doar unul singur, odată ce poate fi folosit succesiv de mai multe ori cu input-urile potrivite. Pentru eficienţă, majoritatea computerelor modeme se bazează pe o paradigmă în care calculele sunt divizate în paşi temporali în cursul cărora informaţia e plimbată între modulele de memorie şi modulele de calcul. Această arhitectură computaţională a fost elaborată între 193 5 şi 1945 de Alan Turing, Konrad Zuse, John Mauchly şi John von Neumann. Mai exact, memoria calculatorului stochează atât datele, cât şi software-ul (un program, adică o listă de instrucţiuni pentru ce are de făcut cu datele). La fiecare pas, o unitate centrală de procesare (CPU) execută următoarea instrucţiune din program, care specifică aplicarea unei funcţii simple asupra anumitor date. Partea din calculator care urmăreşte ce e de făcut în continuare e doar o altă zonă din memoria lui, numită *contorul de program,* care stochează numărul liniei curente din program. Pentru a sări la altă linie din program, copiezi pur şi simplu numărul liniei în contorul de program – acesta e felul în care aşa-numitele instrucţiuni şi bucle „if“ sunt executate.

Calculatoarele de azi câştigă în viteză prin *procesare paralelă,* care reutilizează ingenios modulele: dacă un calcul poate fi des făcut în părţi ce pot fi efectuate în paralel (fiindcă input-ul unei părţi nu presupune output-ul

alteia), atunci părţile pot fi calculate simultan de diferite componente ale hardware-ului.

Calculatorul paralel ultim este *computerul cuantic.* David Deutsch, unul dintre pionierii calculului cuantic, susţine controversata idee conform căreia „computerele cuantice împart informaţia cu un număr uriaş de versiuni ale lor însele din multivers“ şi pot căpăta răspunsuri mai repede aici, în universul nostru, primind într-un anumit sens ajutor de la aceste alte versiuni.[4] Nu ştim încă dacă un computer cuantic competitiv comercial poate fi construit în următoarele decenii – asta depinde de corectitudinea felului în care înţelegem că funcţionează mecanica cuantică şi de capacitatea noastră de a depăşi imense dificultăţi tehnice-, dar companii şi guverne din întreaga lume pariază anual zeci de milioane de dolari pe această posibilitate. Deşi computerele cuantice nu pot accelera calculele obişnuite, au fost elaboraţi algoritmi ingenioşi care pot accelera anumite tipuri de calcule cum ar fi spargerea codurilor şi antrenarea reţelelor neuronale. Un computer cuantic ar putea de asemenea simula eficient comportamentul sistemelor cuantice, între care atomi, molecule şi materiale noi, înlocuind măsurătorile din laboratoarele de chimie, aşa cum simulările pe computere tradiţionale au înlocuit măsurătorile din tunelurile aerodinamice.

CE ESTE ÎNVĂŢAREA:

Deşi un calculator de buzunar mă poate surclasa la un concurs de socoteli, el nu-şi va îmbunătăţi niciodată viteza sau precizia, oricât ar exersa. El nu învaţă: de pildă, de fiecare dată când apăs butonul „radical“, calculează exact aceeaşi funcţie în exact acelaşi fel. în mod asemănător, primul program de calculator care m-a bătut la şah n-a

învăţat niciodată din greşelile lui, ci pur şi simplu a executat funcţia pe care un programator ingenios o proiectase pentru a calcula o mutare bună. Dimpotrivă, când Magnus Carlsen a pierdut primul joc de şah la vârsta de cinci ani, a început un proces de învăţare care a făcut din el, optsprezece ani mai târziu, campion mondial la şah.

Capacitatea de a învăţa este pesemne cel mai fascinant aspect al inteligenţei generale. Am văzut deja cum o bucată de materie neînsufleţită poate memora şi calcula, dar cum poate ea învăţa? Am văzut că găsirea răspunsului la o întrebare dificilă e echivalentă cu calculul unei funcţii, şi că materia aranjată corespunzător poate calcula orice funcţie calculabila. Când noi, oamenii, am construit primele calculatoare de buzunar şi primele programe de şah, *noi* am făcut aranjarea. Pentru ca materia să înveţe, trebuie să se rearanjeze pe ea însăşi ca să se perfecţioneze tot mai mult în calculul funcţiei dorite – supunându-se pur şi simplu legilor fizicii.

Pentru a demistifica procesul învăţării, să vedem cum sisteme fizice foarte simple pot învăţa cifrele lui rt şi ale altor numere. Am văzut mai sus că o suprafaţă cu mai multe văi (figura 2.3) poate fi folosită ca dispozitiv de memorie: de pildă, dacă fundul uneia dintre văi se afla la poziţia $x = n * 3{,}14159$ şi nu există alte văi în apropiere, atunci poţi pune bila la $x = 3$ şi vei vedea cum sistemul calculează zecimalele lipsă lăsând bila să se rostogolească. Să presupunem acum că suprafaţa e alcătuită dintr-un lut moale, iar la început e perfect plată. Dacă vreun pasionat de matematică aşază în mod repetat bila în poziţiile numerelor lui preferate, treptat gravitaţia va crea văi în acele locuri, după care suprafaţa de lut poate fi folosită pentru a regăsi aceste amintiri stocate. Cu alte cuvinte, suprafaţa de lut a învăţat să calculeze cifrele unor numere precum n.

Alte sisteme fizice, cum sunt creierele, pot învăţa mult

mai eficient pe baza aceleiaşi idei. John Hopfield a arătat că reţeaua Iui de neuroni interconectaţi menţionată mai sus poate învăţa într-un mod asemănător dacă o pui în repetate rânduri în anumite stări, treptat va învăţa aceste stări şi se va întoarce la ele din orice stare învecinată. Dacă i-ai văzut pe toţi membrii familiei tale de multe ori, atunci amintirea felului în care arată poate fi declanşată de orice se leagă de ei.

Reţelele neuronale au transformat acum atât biologia, cât şi inteligenţa artificială, iar recent au început să domine un subdo- meniu IA numit *învăţarea automată* (studiul algoritmilor care se perfecţionează prin experienţă). înainte de a înţelege cum pot învăţa asemenea reţele, să vedem cum pot ele calcula. O reţea neuronală e pur şi simplu un grup de neuroni interconectaţi care îşi pot influenţa reciproc comportamentul. Creierul conţine vreo sută de miliarde de neuroni, cam câte stele sunt în galaxia noastră, în medie, fiecare neuron e conectat cu aproximativ o mie de alţi neuroni prin legături numite *sinapse,* iar tăriile acestor circa o sută de bilioane de conexiuni sinaptice este cea care codează majoritatea informaţiei din creier.

Putem desena schematic o reţea neuronală ca pe un ansamblu de puncte reprezentând neuronii, legate prin linii reprezentând sinapsele (figura 2.9). în realitate, neuronii sunt dispozitive electrochimice foarte complicate, constituite din elemente cum sunt axonii şi dendritele. Există diferite tipuri de neuroni, care acţionează în moduri diferite, iar felul în care activitatea unui neuron influenţează alţi neuroni şi timpul de reacţie nu sunt încă pe deplin lămurite. Cercetătorii IA au arătat totuşi că reţelele neuronale pot atinge performanţe de nivel uman în multe sarcini complexe chiar dacă ignorăm toate aceste detalii şi înlocuim neuronii biologici reali cu unii simulaţi extrem de simpli, toţi identici şi supuşi unor

reguli foarte simple. Modelul curent cel mai des folosit pentru o asemenea *reţea neuronală artificială* reprezintă starea unui neuron printr-un singur număr, iar tăria fiecărei sinapse tot printr-un singur număr. în acest model, fiecare neuron îşi actualizează starea la intervale de timp regulate prin medierea input-urilor de la toţi neuronii cu care e conectat, ponderându-le cu tăriile sinaptice, adăugând eventual o constantă, iar apoi aplicând rezultatului ceea ce se numeşte o *Juncţie de activare* pentru a calcula noua stare a neuronului.††† Cel mai simplu mod de a folosi o reţea neuronală ca pe o funcţie este s-o alimentăm cu informaţie dintr-o singură direcţie, ca în figura 2.9, introducând input-ul funcţiei în neuronii din stratul de sus şi extrăgând output-ul de la neuronii din stratul de jos.

Succesul acestor reţele neuronale artificiale simple este încă o dovadă a independenţei de substrat: reţelele neuronale se pare că au o mare putere de calcul independent de detaliile construcţiei lor la nivel elementar. George Cybenko, Kurt Homik, Maxwell Stinchcombe şi Halbert White au demonstrat în 1989 ceva

††† Dacă vă place matematica, două alegeri tipice pentru această funcţie de activare sunt funcţia sigmoidă o(x) = 1/(1 + e- *) şi funcţia rampă a(x) = max [o, x). dar s-a demonstrat că aproape orice funcţie e bună, cu condiţia să nu fie liniară (o dreaptă). Celebrul model al lui Hopfield foloseşte o(x) = -l dacă *x*< o şi a(x) = i dacă x> o. Dacă stările neuronului sunt stocate intr-un vector, atunci reţeaua e actualizată pur şi simplu înmulţind vectorul cu o matrice care stochează cuplajele sinaptice, iar apoi aplică funcţia ofx) tuturor elementelor. *(N. a.)*

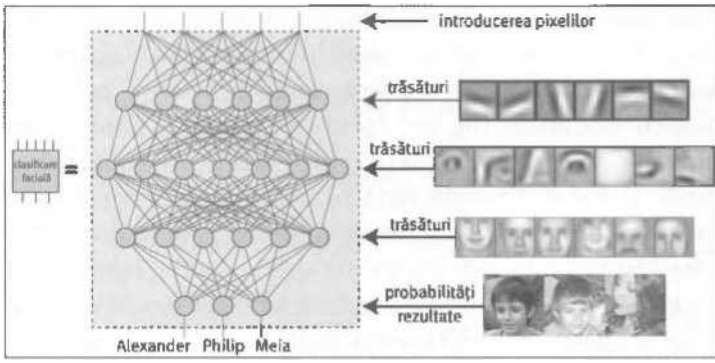


Figura 2.9 O reţea de neuroni poate calcula funcţii la fel ca o reţea de porţî NAND. De exemplu, reţele neuronale artificiale au fost antrenate ca, dacă introduci numere reprezentând strălucirea diferiţilor pixeli dintr-o imagine, să obţii numere reprezentând probabilitatea ca imaginea să înfăţişeze diverşi oameni. Aici, fiecare neuron artificial (cerculeţ) calculează o sumă ponderată a numerelor trimise către el prin conexiuni (linii) de sus, aplică o funcţie simplă ţi transmite rezultatulîn jos, fiecare strat succesiv calculând trăsături de nivel mai înalt. Reţelele tipice de recunoaştere facială conţin sute de mii de neuroni, darîn figură apar doar câţiva pentru a simplifica descrierea.

remarcabil: asemenea reţele neuronale simple sunt universale în sensul că pot calcula *orice* funcţie cu o precizie oricât de mare prin ajustarea corespunzătoare a acelor numere care dau tăria sinapsei. Cu alte cuvinte, probabil că evoluţia n-a făcut neuronii noştri biologici atât de complicaţi pentru că era necesar, ci pentru că era mai eficient – şi pentru că evoluţia, spre deosebire de inginerii umani, nu răsplăteşte proiectele simple şi uşor de înţeles.

Când am aflat asta, m-a mirat cum ceva atât de simplu poate calcula ceva arbitrar de complicat. De pildă, cum poţi calcula până şi ceva atât de banal ca o înmulţire, când tot ce ţi se permite este să calculezi sume ponderate şi să aplici o singură funcţie fixată? Dacă vreţi să vă faceţi o idee, figura 2.10 prezintă felul în care cinci neuroni pot înmulţi două numere arbitrare, iar un singur neuron poate înmulţi trei biţi.

Deşi se poate demonstra că *în teorie* poţi calcula orice

cu o reţea neuronală arbitrar de mare, demonstraţia nu spune dacă o poţi

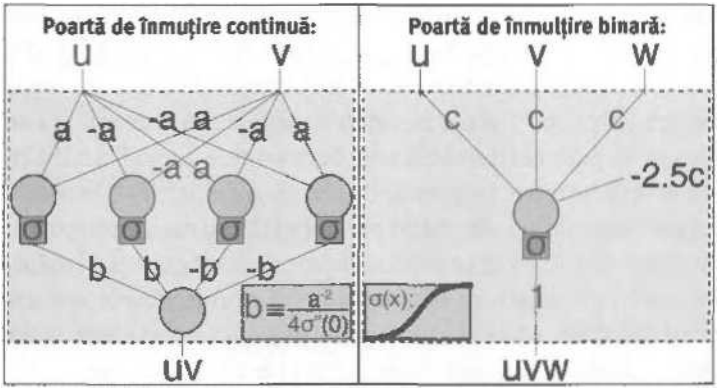


Figura 2.10 Cum poate calcula materia nu folosind porţi NAND ca în figura 2.7, ci neuroni. Fără a intra în detalii, ideea esenţială este că neuronii (artificiali sau biologici) pot face calcule, iar pentru înmulţire e nevoie de mai puţini neuroni decât de porţi NAND. *Detalii opţionale pentru pasionaţii de mate- motică:* Cerculeţele efectuează adunări, pătratele aplică funcţia cr, iar liniile înmulţesc cu constantele marcate pe ele. Input-urile sunt numere reale (stânga) ţi biţi (dreapta). înmulţirea devine arbitrar de precisă când *a* -> o (stânga) şi c-><n (dreapta). Reţeaua din stânga funcţionează pentru orice funcţie <r(x) care e curbată în origine (cu derivata a doua cr"(o) * o), ceea ce se poate demonstra prin dezvoltarea Taylor a lui a». Reţeaua din dreapta presupune ca funcţia a(x) să se apropie de o ţi 1 atunci când x devine foarte mic, respectiv foarte mare, ceea ce se vede observând că *uvw -/* doar dacă u + v + iv = 3. (Exemplele sunt dintr-un articol pe care l-am scris împreună cu studenţii mei Henry Lin ţi David Rolnik, „Why Does Deep and Cheap Learning WorkSo Well?", care poate fi găsit la http://arxiv.0rg/abs/1608.08225.) Combinând numeroaseînmulţiri ţi adunări, poţi calcula orice polinoame, despre care se ştie că pot aproxima orice funcţie netedă.

face *în practică,* având o reţea de dimensiuni rezonabile. Cu cât mă gândesc mai mult, cu atât sunt mai uimit că reţelele neuronale funcţionează atât de bine.

De exemplu, să presupunem că vrem să clasificăm imagini de un megapixel în nuanţe de gri în două categorii, pisici şi câini. Dacă fiecare pixel din acel milion

poate lua, să zicem, 256 de valori, atunci există $256^{iqtX)00<:>}$ imagini posibile, iar pentru fiecare vrem să calculăm probabilitatea de a reprezenta o pisică. Asta

înseamnă că o funcţie arbitrară care are ca input o imagine şi ca output o probabilitate e definită printr-o lista de $256^{1000000}$ probabilităţi, adică mai multe numere decât atomii din universul nostru (circa 10^{78}). Şi totuşi, reţele neuronale cu doar mii sau milioane de parametri îndeplinesc bine sarcina acestei clasificări. Cum reuşesc reţele neuronale de succes să fie „ieftine“, în sensul că au nevoie de atât de puţini parametri? La urma urmei, poţi demonstra că o reţea neuronală suficient de mică pentru a încăpea în universul nostru va eşua lamentabil să aproximeze aproape toate funcţiile, reuşind să îndeplinească doar o parte infimă din toate sarcinile computaţionale pe care i le-ai putea atribui.

M-am gândit mult la asta şi Ia probleme înrudite împreună cu studentul meu Henry Lin. Spre norocul meu, am avut ocazia să colaborez cu studenţi minunaţi, iar Henry este unul dintre ei. Când a intrat prima dată în biroul meu şi m-a întrebat dacă vreau să lucrez cu el, mi-am zis că mai curând eu ar trebui să-l întreb daca îl interesează să lucreze cu mine. Acest băiat modest, prietenos, cu ochi strălucitori, din Shreveport, Louisiana, scrisese deja opt articole ştiinţifice, câştigase Premiul Forbes „so-Under-jo“ şi ţinuse o conferinţă TED cu peste un milion de vizualizări – şi avea doar douăzeci de ani! Un an mai târziu, am scris împreună un articol cu o concluzie surprinzătoare: matematica nu poate răspunde singură la întrebarea de ce funcţionează atât de bine reţelele neuronale, este nevoie şi de fizică. Am descoperit că acea clasă de funcţii pe care ni le oferă fizica şi asupra cărora ne trezeşte interesul pentru a le calcula este de asemenea o clasă extrem de redusă pentru că, din motive pe care încă nu le înţelegem pe deplin, legile fizicii sunt remarcabil de simple. Mai mult, infima parte a funcţiilor pe care reţelele neuronale le pot calcula e foarte asemănătoare cu infima parte asupra căreia fizica ne trezeşte

interesul! Am arătat apoi că reţelele neuronale de învăţare profunde (sunt numite „profunde“ dacă au în compoziţie multe straturi) sunt mult mai eficiente decât cele subţiri pentru multe dintre aceste funcţii de interes. De pildă, împreună cu David Rolnik, un alt student excepţional de Ia MIT, am demonstrat că sarcina simplă de a înmulţi *n* numere necesită 2^n neuroni într-o reţea cu un singur strat, dar numai circa 4» neuroni într-o reţea profundă. Asta explică nu doar de ce reţelele neuronale sunt la modă printre cercetătorii IA, ci şi de ce am dezvoltat în creierele noastre reţele neuronale: dacă am dezvoltat creiere pentru a prezice viitorul, atunci e logic să fi dezvoltat o arhitectură computaţională adaptată tocmai acelor probleme de calcul care contează în lumea fizică.

După ce am văzut cum funcţionează şi calculează reţelele neuronale, să ne întoarcem la problema învăţării. Mai precis, cum îşi poate îmbunătăţi capacitatea de calcul o reţea neuronală actua- lizându-şi sinapsele?

în celebra sa carte din 1949 *The Organization ofBehavior:A Neuropsychological Theory,* psihologul canadian Donald Hebb susţinea că dacă doi neuroni învecinaţi se activează mereu în acelaşi timp, cuplajul lor sinaptic se întăreşte, aşa încât ei învaţă să se activeze reciproc – idee surprinsă de expresia „Fire together, wire together“ („Activaţi împreună, cablaţi împreună“). Deşi detaliile felului în care învaţă creierul sunt încă departe de a fi înţelese, iar cercetările arată că în numeroase cazuri răspunsurile sunt mult mai complicate, s a demonstrat că până şi această regulă simplă de învăţare (numită învăţare hebbiană) permite reţelelor neuronale să înveţe lucruri interesante. John Hopfield a arătat că învăţarea hebbiană permite reţelei lui neuronale ultrasimplificate să stocheze multe amintiri complexe doar prin expunerea repetată la acestea. Asemenea

expunere la informaţie pentru a învăţa din ea se numeşte de regulă „antrenament“ când se referă la reţele neuronale artificiale (ori la animale sau oameni care capătă aptitudini), deşi mai nimerit ar fi „studiu“, „educaţie“ sau „experienţă“. Reţelele neuronale artificiale din sistemele LA de azi tind să înlocuiască învăţarea hebbiană cu reguli de învăţare mai sofisticate, purtând denumiri întortocheate precum „retropropagare“ *(„back-propagation“)* şi „coborâre stocastică în gradient“ *(„stochastic gradient descenf),* însă ideea de bază e aceeaşi: există o regulă deterministă simplă, asemănătoare unei legi a fizicii, prin care sinapsele sunt actualizate din timp în timp. Şi, ca prin magie, această regulă simplă poate face ca reţelele neuronale să înveţe calcule extrem de complexe dacă antrenamentul se desfăşoară cu cantităţi mari de date. Nu ştim încă exact ce reguli de învăţare foloseşte creierul nostru, dar indiferent care ar fi răspunsul, nu avem vreun indiciu că ele ar încălca legile fizicii.

După cum majoritatea computerelor digitale câştigă în eficienţă împărţindu-şi munca în mai mulţi paşi şi refolosind repetat modulele de calcul, la fel fac şi multe reţele neuronale artificiale şi biologice. Greierele conţin părţi pe care informaticienii le numesc reţele *recurente* (spre deosebire de reţelele *feedforward),* în care informaţia poate curge în mai multe direcţii (nu într-una singură), aşa încât output-ul curent poate deveni apoi input. Reţeaua de porţi logice din microprocesorul unui laptop este de asemenea recurentă, în următorul sens: îşi refoloseşte constant informaţia din trecut şi lasă noua informaţie de input (de la tastatură, trackpad, cameră video etc.) să influenţeze calculul curent, care, la rândul lui, determină informaţia de output către ecran, difuzor, imprimantă sau reţea wireless. în mod asemănător, reţeaua de neuroni din creier e recurentă, lăsând

informaţia de input provenind de la ochi, urechi şi alte organe senzoriale să influenţeze calculul curent, care, la rândul lui, determină informaţia de output către muşchi.

Istoria învăţării e cel puţin la fel de lungă ca însăşi istoria vieţii, căci orice organism autoreproducător copiază şi procesează informaţia – comportament care a fost cumva învăţat. în epoca Vieţii 1.0 însă, organismele n-au învăţat în cursul vieţii lor regulile de procesare a informaţiei şi de reacţie erau determinate de ADN-ul lor moştenit, aşa încât învăţarea avea loc lent, la nivelul speciei, prin evoluţie darwinistă de-a lungul generaţiilor.

Cu aproximativ o jumătate de miliard de ani în urmă, anumite structuri genetice au găsit o modalitate de a face ca animale conţinând reţele neuronale să poată învăţa comportamente din experienţele trăite în cursul vieţii. Apăruse Viaţa z.o, şi datorită capacităţii ei de a învăţa incomparabil mai repede, surclasându-şi competitorii, s-a răspândit fulgerător pe glob. După cum am văzut în capitolul i, viaţa a progresat în privinţa învăţării într-un ritm tot mai rapid. 0 specie semănând cu maimuţele antropoide a căpătat un creier atât de bine adaptat la dobândirea cunoştinţelor, încât a învăţat să folosească unelte, să aprindă focul, să vorbească şi să creeze o societate globală complexă. Societatea însăşi poate fi privită ca un sistem care memorează, calculează şi învaţă, toate acestea într-un ritm accelerat, căci o invenţie o face posibilă pe următoarea: scris, presă tipăritä, ştiinţă modernă, calculatoare, intemet etc. Ce vor nota istoricii ştiinţei din viitor în continuarea listei invenţiilor? Inteligenţă artificială, cred eu.

Progresele explozive ale memoriei şi puterii de calcul ale computerelor (figurile 2.4 şi 2.8) au determinat un progres spectaculos al IA, dar a trecut mult timp până ca *învăţarea* automată să se maturizeze. Când calculatorul IBM Deep Blue l-a învins la şah pe campionul mondial

Gări Kasparov, în 1997, atuurile lui ţineau de memorie şi calcul, nu de învăţare. Inteligenţa lui computaţională fusese creată de o echipă de oameni, iar Deep Blue îşi învingea creatorii graţie vitezei lui mai mari de calcul, care îi permitea să analizeze mai multe posibilităţi. Când calculatorul IBM Watson l-a învins pe campionul emisiunii-concurs *Jeopardy!,* nici acesta nu s-a bazat pe învăţare, ci pe aptitudini specializate programate şi pe memorie şi viteză superioare. Acelaşi lucru e valabil pentru majoritatea primelor realizări din robotică, de la locomo- ţia cu picioare la automobile care se conduc singure şi rachete care aterizează singure.

Dimpotrivă, forţa motrice a celor mai recente realizări IA a fost *învăţarea* automată. Priviţi, de exemplu, figura 2.11. E uşor pentru noi să spunem ce reprezintă fotografia, dar ca să programezi o funcţie pentru care input-ul să fie doar culorile tuturor pixelilor unei imagini, iar output-ul să fie o descriere precisă cum ar fi „Un grup de copii jucându-se cu un disc zburător", i-a pus în dificultate pe toţi cercetătorii IA din lume vreme de decenii. Şi totuşi o echipă de la Google a reuşit asta în 2014.[5] Dacă introduci alt set de pixeli coloraţi, ţi se răspunde „O turmă de elefanţi umblând printr-un câmp cu iarbă uscată" – iarăşi corect. Cum a fost posibil? Prin programarea artizanală a algoritmilor, în stil Deep Blue,

Figura 2.11 „Un grup de copii jucându-se cu un disc zburător“ – aceasta e descrierea dată de un calculator care nu înţelege nimic despre copii, jocuri sau discuri zburătoare.

pentru a detecta discuri zburătoare, chipuri etc.? Nu, ci prin crearea unei reţele neuronale relativ simple, fără cunoştinţe despre lumea fizică sau conţinutul ei, şi lăsând-o apoi să înveţe prin expunerea Ia cantităţi imense de date. Jeff Hawkins scria în 2004 că „nici un calculator nu poate [...] vedea la fel de bine ca un şoarece“, dar acele zile sunt de mult apuse.

Aşa cum nu înţelegem pe deplin cum învaţă copiii noştri, încă nu înţelegem pe deplin cum învaţă aceste reţele neuronale, şi de ce uneori dau greş. E clar însă că ele sunt deja extrem de utile şi declanşează un val de investiţii în învăţarea profundă. învăţarea profundă a transformat multe aspecte ale „vederii” calculatorului, de Ia transcrierea scrisului de mână la analiza video în timp real a automobilelor care se conduc singure. A transformat de asemenea capacitatea calculatoarelor de a transforma vorbirea în text şi a o traduce în alte limbi, chiar în timp real – de aceea putem acum vorbi cu asistenţi personali digitali cum sunt Siri, Google Now şi Cortana. Acele obositoare teste „CAPTCHA“, ca să

convingă site-ul că suntem oameni, devin tot mai dificile, pentru a depăşi posibilităţile tehnologiei învăţării profunde. în 2015, Gogle DeepMind <1 lansat un sistem IA folosind învăţarea profundă care se putea descurca în zeci de jocuri computerizate la fel cum ar face-o un copil – fără să aibă instrucţiuni, iar curând a învăţat să joace mai bine decât orice om. în 2016, aceeaşi companie a construit AlphaGo, un sistem pentru jocul de go, folosind învăţarea profundă ca să evalueze diferite poziţii, care l-a învins pe campionul mondial Ia go. Acest progres stimulează un cerc virtuos, aducând şi mai multe finanţări şi talente în cercetarea IA, care generează progresul ulterior.

Am examinat în acest capitol natura inteligenţei şi dezvoltarea ei până acum. Cât va mai dura până ca maşinile să ne depăşească în toate sarcinile cognitive? Nu ştim şi trebuie să acceptăm posibilitatea ca răspunsul să fie „niciodată”. Mesajul fundamental al acestui capitol este însă că trebuie să luăm în considerare şi posibilitatea că se *va* întâmpla, poate chiar în cursul vieţii noastre. La urma urmei, materia poate fi aranjată aşa încât, supunându-se legilor fizicii, să poată memora, calcula şi învăţa – şi nu-i nevoie ca materia să fie biologică. Cercetătorii IA au fost adesea acuzaţi că promit prea mult şi realizează prea puţin, dar adevărul e că unii dintre criticii lor n-au un palmares mai grozav. Unii împing pur şi simplu obiectivele mai departe, definind inteligenţa ca pe ceea ce calculatoarele încă nu pot face, sau ca pe ceea ce ne impresionează. Maşinile sunt acum bune sau excelente la calcule, şah, demonstrarea teoremelor matematice, analiză bursieră, descrierea imaginilor, şofat, jocuri video, go, sinteză vocală, transcrierea vorbirii, traducere şi diagnosticarea cancerului, însă unii critici vor strâmba din nas: „Mă rog, dar asta nu-i inteligenţă acfeyărată!” Ei ar putea spune că inteligenţa adevărată presupune doar crestele munţilor din peisajul lui Moravec

(figura 2.2) care n-au fost încă inundate, la fel cum se spunea înainte că ar trebui să conteze descrierea imaginilor şi jocul de go – iar nivelul apei creştea mereu.

Presupunând că nivelul apei va continua să crească, impactul IA asupra societăţii va fi tot mai mare. Cu mult înainte ca IA să ajungă la nivel uman în toate sarcinile, vom fi confruntaţi cu alegeri şi provocări fascinante legate de erori, legi, arme şi joburi. Care sunt acestea şi cum ne putem pregăti cât mai bine pentru a le face faţă – iată subiectul următorului capitol.

REZUMAT:

• Inteligenţa, definită drept capacitatea de a îndeplini obiective complexe, nu poate fi măsurată cu un singur IQ, cu un singur spectru de capacităţi pentru toate obiectivele.

• Inteligenţa artificială de azi este de regulă *îngustă,* fiecare sistem putând îndeplini numai obiective foarte precise, în vreme ce inteligenţa umană are un spectru remarcabil de larg.

• Memoria, calculul, învăţarea şi inteligenţa sugerează ceva abstract, impalpabil şi eteric, fiindcă sunt *independente de substrat:* au o viaţă proprie care nu depinde de detaliile substratului lor material şi nu le reflectă.

• Orice bucată de materie poate fi substratul *memoriei* dacă are mai multe stări stabile.

• Orice materie poate fi *computroniu,* substratul calculului, dacă e compusă din elemente ce pot fi combinate pentru a executa orice funcţie. Porţile NAND şi neuronii sunt două exemple importante de asemenea „atomi computaţionali“ universali.

• O reţea neuronală e un substrat puternic pentru *învăţare* fiindcă, supu- nându-se pur şi simplu legilor fizicii, se poate rearanja pentru a se auto- perfecţiona în efectuarea calculelor dorite.

• Datorită simplităţii frapante a legilor fizicii, pe noi, oamenii, ne preocupă doar o infimă parte din toate problemele computaţionale Imaginabile, iar reţelele neuronale sunt de regulă remarcabil de pricepute tocmai la rezolvarea acestei părţi minuscule.

• Odată ce tehnologia îşi dublează puterea, adesea poate fi folosită pentru a proiecta şi construi o tehnologie (a rândul ei de două ori mai puternică, declanşând dublarea repetată în spiritul legii lui Moore. Timp de peste un secol, preţul tehnologiei informaţiei s-a înjumătăţit cam la fiecare doi ani, făcând cu putinţă apariţia epocii informaţiei.

• Dacă progresul IA continuă, cu mult înainte ca IA să atingă nivelul uman la toate capitolele vom fi confruntaţi cu alegeri şi provocări fascinante legate de erori, legi, arme şi joburî – pe care le vom examina în următorul capitol.

CAPITOLUL 3

Viitorul apropiat: progrese, erori, legi, arme şi joburi

> *Dacă nu schimbăm repede direcţia, ne îndreptăm sprepieire.* — Irwin Corey

Ce înseamnă să fii om în ziua şi în epoca de azi? De pildă, ce preţuim cu adevărat la noi, ce ne deosebeşte de alte forme de viaţă şi de maşini? Ce preţuiesc la noi alţi oameni, aşa încât unii vor să ne ofere joburi? Indiferent de răspunsurile noastre la un moment dat, e limpede că ascensiunea tehnologiei le modifică treptat.

Iată, de pildă, cazul meu. Ca om de ştiinţă, mă mândresc că îmi stabilesc singur obiectivele, îmi pun la lucru creativitatea şi intuiţia pentru a aborda o gamă largă de probleme nerezolvate şi folosesc limbajul pentru a împărtăşi altora ceea ce descopăr. Norocul meu – societatea e dispusă să mă plătească pentru jobul ăsta, în schimb, cu secole în urmă, la fel ca mulţi alţii, aş fi putut să-mi construiesc personalitatea în jurul meseriei de fermier sau de meşteşugar, dar ascensiunea tehnologiei a redus între timp asemenea profesiuni la o parte minusculă din forţa de muncă. Asta înseamnă că nu oricine îşi mai poate construi identitatea în jurul muncilor agricole sau meşteşugurilor.

Pe mine unul nu mă deranjează că maşinile de azi mă depăşesc în aptitudini manuale precum săpatul şi împletitul, fiindcă acestea nu sunt nici hobby-urile mele,

nici sursele mele de venit sau de satisfacţie personală. Orice iluzii mi-aş fi făcut în privinţa unor asemenea talente, ele mi-au fost spulberate la vârsta de opt ani, când şcoala m-a obligat să urmez lecţii de împletit, şi n-aş fi luat notă de trecere dacă nu m-ar fi ajutat un binevoitor din clasa a cincea, căruia i s-a făcut milă de mine.

Cum însă tehnologia continuă să se perfecţioneze, ascensiunea IA va eclipsa oare în cele din urmă şi acele aptitudini care îmi conferă acum stima de sine şi valoarea pe piaţa muncii? Stuart Russell mi-a mărturisit că el şi mulţi dintre colegii lui cercetători IA trecuseră recent printr-un moment de uluială când au văzut că IA face un lucru Ia care nu se aşteptau. în acelaşi spirit, daţi-mi voie să vă vorbesc despre câteva dintre momentele mele de uluială, care cred că vestesc depăşirea în curând a capacităţilor umane.

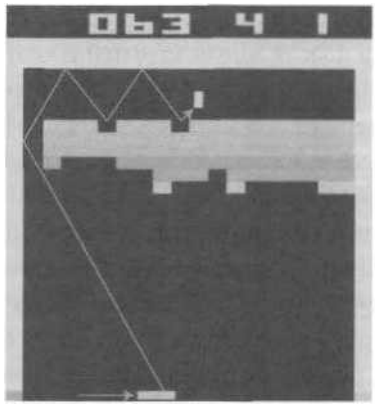
PROGRESE

Agenţi de consolidare a învăţării profunde

în 2014 am rămas perplex când am văzut un video în care un sistem IA DeepMind învăţa să joace jocuri pe calculator. Mai exact, IA juca Breakout (vezi figura 3.1), un joc Atari clasic pe care mi-l aminteam din adolescenţă. Obiectivul era să manevrezi o paletă aşa încât să faci ca o minge să ricoşeze în mod repetat pe un zid de cărămizi; de fiecare dată când loveai o cărămidă, ea dispărea şi scorul tău creştea.

Scrisesem şi eu pe vremuri câteva jocuri pentru calculator, şi ştiam că nu era greu să scrii un program care să joace Breakout – dar nu asta făcuse echipa DeepMind. Ei creaseră în schimb o IA ca o *tabula rasa,* care nu ştia nimic despre acest joc – sau despre vreun altul, sau despre *concepte* cum sunt jocurile, paletele,

cărămizile ori mingile. Tot ce primea IA la intervale regulate era o lungă listă de numere: scorul curent şi lista numerelor în care noi (dar nu şi IA) am recunoaşte felul în care erau colorate diferite părţi ale ecranului. Tot ce i se spusese era să maximizeze scorul producând la intervale regulate numere în care noi (dar nu şi IA) am recunoaşte coduri pentru tasta care trebuie apăsată.



A gura 3,2 După ce a învăţat să joace Breakout pornind de la zero, folosind învăţarea consolidată pentru a maximiza scorul, IA DeepMind a descoperit strategia optimă-, să facă o gaură în extremitatea din stânga a zidului de cărămizi ţi să (ase mingea să tot ricoşeze în spatele lui, adunând rapid puncte. Săgeţile reprezintă traiectoriile din trecut ale mingii ţi paletei.

La început, ÎA juca îngrozitor: mişca paleta înainte şi înapoi aparent la întâmplare, ratând mai mereu mingea. După un timp, a părut să-şi dea seama că mişcarea paletei spre minge era o idee bună, dar încă rata de cele mai multe ori. Se perfecţiona însă pe măsură ce exersa, devenind curând mai bună decât fusesem eu vreodată, retumând fără greş mingea oricât de rapid ar fi venit spre ea. Iar apoi am rămas perplex; a descoperit strategia uimitoare de maximizare a scorului, ţintind mereu colţul din stân- ga-sus pentru a face o gaură în zid şi a bloca mingea între partea de sus a zidului şi bariera din spatele

lui. Era într-adevăr un lucru inteligent. Demis Hassabis avea să-mi spună că programatorii din acea echipă DeepMind nu cunoşteau acest truc, şi l-au învăţat de la IA pe care o construiseră. Vă recomand să priviţi un document video la link-ul pe care vi-l pun la dispoziţie.[1]

Era aici o trăsătură umană care m-a tulburat: priveam o IA care avea un obiectiv, iar în cele din urmă a învăţat să-l îndeplinească mai bine decât creatorii ei. în capitolul precedent am definit inteligenţa drept capacitatea de a îndeplini obiective complexe, deci
în acest sens IA a lui DeepMind devenea tot mai inteligentă în faţa ochilor mei (deşi doar în sensul îngust al acestui joc particular), în primul capitol am întâlnit ceea ce informaticienii numesc *agenţi inteligenţh* entităţi care adună informaţie de la senzori despre mediul lor, iar apoi procesează informaţia pentru a hotărî cum să reacţioneze asupra mediului. Deşi IA a lui DeepMind pentru acest joc trăia într-o lume virtuală extrem de simplă, alcătuită din cărămizi, palete şi mingi, nu puteam nega că era un agent inteligent.

Cei de la DeepMind şi-au publicat curând metoda şi şi-au dezvăluit codul, explicând că folosea ideea foarte simplă, dar puternică, de învăţare consolidată profundă *{deep reinforcement leaming).*[2] învăţarea consolidată elementară e o tehnică clasică de învăţare automată, inspirată din psihologia behavioristă, în care primirea unei recompense sporeşte tendinţa de a repeta ceva, şi viceversa. La fel cum un câine învaţă să facă giumbuşlucuri când asta sporeşte probabilitatea de a primi imediat încurajări sau hrană de Ia stăpân, IA a lui DeepMind a învăţat să mişte paleta pentru a prinde mingea, fiindcă asta sporea probabilitatea de a primi imediat mai multe puncte. DeepMind a combinat această idee cu învăţarea profundă: a fost antrenată o reţea neuronală profundă, Ia fel ca în capitolul precedent, pentru a prezice

câte puncte vor fi câştigate în medie apăsând fiecare tastă permisă a claviaturii, iar apoi IA a selectat tasta pe care reţeaua neuronală a găsit-o cea mai promiţătoare, dată fiind starea curentă a jocului. în lista trăsăturilor cu care mă mândresc ca om am inclus capacitatea de a aborda o gamă largă de probleme nerezolvate. Să poţi juca Breakout şi să nu fii în stare de nimic altceva denotă o inteligenţă foarte îngustă. Pentru mine, ce este cu adevărat important în rezultatul obţinut de echipa DeepMind e faptul că învăţarea consolidată profundă reprezintă o tehnică absolut generală. Ei au pus exact aceeaşi LA să exerseze alte 49 de jocuri Atari, iar aceasta a învăţat să-şi depăşească experimentatorii umani în 29 dintre ele, de Ia Pong la Boxing, Video Pinball şi Space Invaders.

N-a durat mult până când aceeaşi idee de IA să fie testată în jocuri mai modeme, cu lumi tridimensionale, nu doar bidimensioinale.

Curând, competitorii din San Francisco ai lui DeepMind în campi- onattd IA au lansat o platformă numită Universe, unde DeepMind şi alţi agenţi inteligenţă pot exersa interacţionând cu un întreg computer ca şi cum ar fi un joc: apăsând pe orice, tastând orice, deschizând şi rulând orice software accesibil- intrând pe un browser şi navigând online, de pildă.

Privind spre viitorul învăţării consolidate profunde şi al perfecţionării ei, nu se întrezăreşte capătul. Potenţialul ei nu se mărgineşte Ia lumile jocurilor virtuale, căci dacă ai fi robot ai putea privi viaţa însăşi ca pe un joc. Stuart Russell mi-a spus că primul lui mare moment de perplexitate a fost când a văzut robotul Big Dog alergând pe o pantă acoperită cu zăpadă şi rezolvând elegant problema locomoţiei cu picioare Ia care el muncise ani de-a rândul ca s-o rezolve.3 Iar când în fine obiectivul a fost atins, în 2008, acesta a necesitat o cantitate imensă de

muncă a unor programatori iscusiţi. După succesul celor de la DeepMind, nu există motive pentru care un robot să nu poată folosi o variantă a învăţării consolidate profunde ca să înveţe singur să meargă, fără ajutorul unor programatori umani: nu-i nevoie decât de un sistem care să-i dea puncte atunci când face progrese. Roboţii din lumea reală ar putea şi să înveţe să înoate, să zboare, să joace ping-pong, să se lupte şi să îndeplinească o listă aproape nesfârşită de alte sarcini motorii, fără ajutor din partea programatorilor umani. Pentru a accelera lucrurile şi a reduce riscul de a se bloca sau deteriora în cursul procesului de învăţare, ar trebui probabil să parcurgă primele etape ale învăţării în realitatea virtuală.

Intuiţie, creativitate şi strategie

Alt moment definitoriu pentru mine a fost acela când sistemul IA AlphaGo al lui DeepMind a câştigat meciul de go în cinci jocuri împotriva lui Lee Sedol, considerat cel mai bun jucător din lume la începutul secolului XXI.

Se aştepta ca la un moment dat jucătorii umani de go să fie detronaţi de maşini, fiindcă aceasta se întâmplase cu colegii lor şahişti în urmă cu două decenii. Majoritatea specialiştilor în go preziceau însă că va mai trece un deceniu, aşa încât triumful lui AlphaGo a fost un moment de răscruce pentru ei, ca şi pentru mine. Nick Bostrom şi Ray Kurzweil au subliniat cât de dificil era ca IA să progreseze în acest domeniu, ceea ce transpare şi din interviurile cu Lee Sedol dinainte şi de după ce a pierdut primele trei jocuri:

- octombrie 2015: „Pe baza nivelului constatat (...] cred că voi câştiga clar jocul.“
- februarie 2016: „Am aflat că IA al lui DeepMind de la Google e surprinzător de puternic, şi devine şi mai puternic, dar sunt încrezător că măcar de data asta

pot câştiga.“

■ 9 martie 2016: „Sunt foarte surprins, pentru că nu mi-am închipuit că pot pierde.“

• 10 martie 2016: ‚^m rămas fără cuvinte. [...J Sunt şocat. Recunosc că [...] al treilea joc nu va fi uşor pentru mine.“

• 12 martie 2016: „Mă simt neputincios.“

La un an după ce a jucat cu Lee Sedol, un AlphaGo perfecţionat a jucat cu primii douăzeci de jucători din lume, fără să piardă un singur meci.

De ce era pentru mine un lucru atât de important? Mărturisesc că priveam intuiţia şi creativitatea ca pe două dintre trăsăturile mele intrinsec umane, şi, după cum voi arăta imediat, am simţit că AlphaGo le manifesta pe ambele.

Jucătorii de go aşază pe rând piese albe şi negre pe o tablă de 19x19 (vezi figura 3.2). Există cu mult mai multe poziţii de go posibile decât atomii din universul nostru, ceea ce înseamnă că a încerca să analizezi toate şirurile interesante de mutări viitoare devine rapid inutil. Jucătorii se bazează aşadar din plin pe intuiţia subconştientă pentru a-şi completa judecata conştientă, la experţi dezvoltându-se o stranie senzaţie legată de poziţiile puternice şi de cele slabe. După cum am văzut în capitolul precedent, rezultatele învăţării profunde amintesc uneori de intuiţie: o reţea neuronală profundă poate determina că o imagine reprezintă o pisică, fără să poată explica de ce. Echipa DeepMind a pariat deci pe ideea că învăţarea profundă ar putea să recunoască nu doar pisici, dar şi poziţii puternice de go. Ideea centrală introdusă în AlphaGo a fost de a reuni puterea intuitivă a învăţării profunde cu puterea logică a lui GOFAI – acronimul pentru „Good Old-Fashioned AI“ („destoinica IA demodată“) dinainte de revoluţia învăţării profunde. A fost folosită o uriaşă bază de date a poziţiilor de go din

jocuri dintre oameni, precum şi din jocuri în care AlphaGo se confruntase cu o clonă a lui însuşi, şi a fost antrenată o reţea neuronală profundă pentru a prezice la fiecare poziţie probabilitatea ca albul să câştige în cele din urmă. A fost de asemenea antrenată o reţea pentru a prezice următoarele mutări probabile. Apoi, aceste două reţele au fost combinate cu o metodă GOFAI care căuta în mod ingenios într-o listă simplificată de şiruri ulterioare probabile de mutări pentru a identifica următoarea mutare care ar conduce la poziţia cea mai puternică.

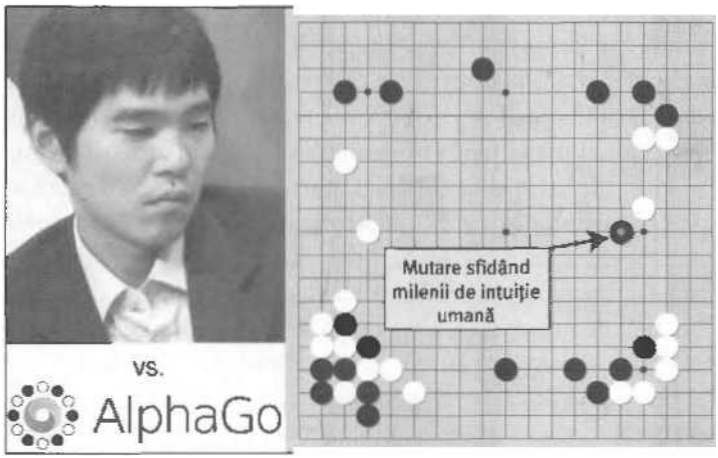


Figura 3.2 IA AlphaGo a lui DeepMind a făcut o mutare extrem de inventivă pe linia 5, sfidând milenii de înţelepciune umană, care peste vreo 50 de mutări s-a dovedit a fi crucială în victoria asupra maestrului go Lee Sedol.

Acest mariaj între intuiţie şi logică a dat naştere unor mutări care nu erau doar puternice, dar şi extrem de inventive. De pildă, milenii de înţelepciune go dictează că la începutul jocului cel mai bine e să muţi pe linia a treia sau a patra de la margine. Dacă muţi pe linia a treia, obţii pe termen scurt câştig teritorial spre marginea tablei, în schimb, dacă muţi pe linia a patra capeţi influenţă strategică pe termen lung spre centru.

La mutarea numărul 37 din al doilea joc, AlphaGo a

şocat lumea go-ului, sfidând o veche înţelepciune şi jucând pe linia a cin- cea (figura 3.2), de parcă ar fi fost mai încrezător decât un om în capacitatea de a plănui pe termen lung, favorizând astfel avantajul strategic în faţa celui pe termen scurt. Comentatorii au fost uluiţi, iar Lee Sedol chiar s-a sculat şi a părăsit încăperea.[4] După vreo 50 de mutări însă, lupta din colţul stânga-jos al tablei s-a încheiat făcându-se legătura cu piesa neagră de la mutarea 37! Asta a condus la câştigarea jocului, întărind ideea că mutarea lui AlphaGo pe linia a cincea a fost una dintre cele mai inovatoare din istoria go-ului.

Datorită aspectelor lui intuitive şi creatoare, go este privit mai curând ca o formă de artă decât ca un joc oarecare. în vechea Chină era considerat una dintre cele patru „arte esenţiale", alături de pictură, caligrafie şi muzica *qin,* şi a rămas extrem de popular în Asia, unde aproape 300 de milioane de oameni au urmărit primul joc între AlphaGo şi Lee Sedol. Ca urmare, lumea go-ului a fost şocată de rezultat şi a găsit că victoria lui AlphaGo e un moment crucial pentru omenire. Ke Jie, liderul clasamentului de go la acea vreme, a declarat: „Omenirea a jucat go de mii de ani, dar, după cum ne-a demonstrat IA, abia am ajuns la suprafaţa lui. [...] Reunirea jucătorilor umani cu cei computerizaţi deschide o nouă eră. (...] împreună, om şi IA pot găsi adevărul jocului go." 0 asemenea colaborare rodnică om-maşină pare promiţătoare în multe domenii, inclusiv în ştiinţă, unde putem spera ca IA să ne ajute să înţelegem mai adânc lucrurile şi să ne dezvoltăm pe deplin potenţialul.

AlphaGo îmi sugerează şi o altă idee importantă pentru viitorul apropiat: combinarea intuiţiei învăţării profunde cu logica GOFAI poate produce *strategii* fără egal. Go fiind jocul strategic absolut, se cuvine acum ca IA să concureze (sau să ajute) cele mai bune strategii umane dincolo de tabla de joc – de pildă în strategiile de

investiţie, politice şi militare. Asemenea probleme strategice din lumea reală sunt de regulă complicate de psihologia umană, lipsa informaţiei şi de factori care trebuie modelaţi aleatoriu, dar sisteme IA pentru jocul de poker au demonstrat deja că nici una dintre aceste provocări nu e insurmontabilă.

Limbajul natural

Alt domeniu în care progresele IA m-au uimit recent este limbajul, în tinereţe mi-a plăcut să călătoresc, iar curiozitatea pentru alte culturi şi limbi joacă un rol important în identitatea mea. Am vorbit de mic suedeză şi engleză, am învăţat germană şi spaniolă la şcoală, portugheză şi română în urma a două căsnicii, şi am învăţat singur ceva rusă, franceză şi mandarină.

Dar IA a ajuns, după o realizare măreaţă din 2016, nu există aproape deloc limbi între care eu pot să traduc mai bine decât sistemul IA dezvoltat de echipamentul creierului lui Google.

Aţi înţeles ce-am spus? Voiam de fapt să spun următoarele:

Dar IA m-a ajuns din urmă, iar după o realizare importantă din 2016, aproape ca n-au rămas limbi între care eu să pot traduce mai bine decât sistemul IA dezvoltat de echipa Google Brain.

Prima dată am tradus fraza în spaniolă, iar pe urmă înapoi în engleză, folosind o aplicaţie instalată pe laptopul meu acum câţiva ani. în 2016, echipa Google Brain şi-a actualizat serviciul gratuit Google Translate pentru a folosi reţele neuronale recu rente profunde, iar îmbunătăţirea faţă de mai vechile sisteme GOFAI a fost spectaculoasă:[6]

Dar IA m-a ajuns din urmă, iar după o realizare din 2016, aproape că n-au rămas limbi care să poată traduce

mai bine între ele decât sistemul IA dezvoltat de echipa Google Brain.

După cum vedeţi, pronumele „eu“ s-a pierdut în ocolul prin spaniolă, ceea ce din păcate a schimbat înţelesul. Pe-aproape, dar nu tocmai! Trebuie totuşi spus, în apărarea IA creată de Google, că adesea sunt criticat fiindcă scriu fraze inutil de lungi, greu de descâlcit, şi am ales una dintre cele mai torturate pentru exemplul meu. Această IA traduce de multe ori impecabil propoziţiile tipice. A făcut aşadar vâlvă când a apărut, şi e suficient de utilă pentru a fi folosită zilnic de milioane de oameni. În plus, graţie progreselor în învăţarea profundă pentru conversiile vorbire-text şi text-vor- bire, aceşti utilizatori pot acum vorbi în telefoanele lor inteligente într-o limbă şi asculta rezultatul tradus.

Procesarea limbajului natural este unul dintre domeniile IA cele mai dinamice, iar eu cred că viitoarele succese vor avea un impact puternic, fiindcă limbajul e o trăsătură umană esenţială. Cu cât se va perfecţiona IA în predicţie lingvistică, cu atât se va pricepe mai bine să compună răspunsuri la e-mailuri sau să continue o conversaţie orală. Asta ar putea da aparenţa, cel puţin unui novice, că se desfăşoară o gândire umană. Sistemele de învăţare profundă urmează deci paşii unui prunc pe calea spre faimosul test Turing, în care maşina trebuie să converseze în scris suficient de bine pentru a face pe cineva să creadă că e un om.

IA pentru procesarea limbajului mai are însă o cale lungă de străbătut. Deşi mărturisesc că mă simt puţin dezamăgit când sunt tradus de o IA, îmi revin când îmi amintesc că ea nu *înţelege deloc* ceea ce spune. Din antrenamentul cu seturi masive de date, ea descoperă tipare şi relaţii implicând cuvinte, fără să raporteze aceste cuvinte la ceva din lumea reală. De exemplu, ar putea reprezenta fiecare cuvânt printr-o listă cu mii de numere

care arată cât de asemănător este cu anumite alte cuvinte. Ar putea conchide de aici că diferenţa dintre „rege“ şi „regină“ e asemănătoare celei dintre „soţ“ şi „soţie“ – dar tot n-are vreun indiciu despre ce înseamnă să fii bărbat sau femeie, ori chiar despre faptul că există o realitate fizică, cu spaţiu, timp şi materie.

Odată ce testul Turing presupune în fond o păcăleală, el a fost criticat pentru că vizează mai curând naivitatea umană decât veritabila inteligenţă artificială. Dimpotrivă, un test concurent numit *Provocarea Schemei Winograd* ţinteşte drept spre înţelegerea de bun-simţ care scapă sistemelor de învăţare profundă actuale. Noi, oamenii, apelăm în mod obişnuit la cunoştinţe despre lumea reală când analizăm o frază pentru a ne da seama la ce anume se referă un pronume. De pildă, o provocare Winograd tipică pune întrebarea despre care „ei“ este vorba aici:

1. „Consilierii municipali au refuzat să le dea demonstranţilor aprobarea, fiindcă ei se temeau de violenţă.“
2. „Consilierii municipali au refuzat să le dea demonstranţilor aprobarea, fiindcă ei pledau pentru violenţă.“

Există o competiţie anuală IA pentru a răspunde la asemenea întrebări, iar rezultatele sunt deocamdată execrabile.[7] Această provocare precisă, a înţelege ce anume se referă la ce, l-a spulberat până şi pe Google Translate când am înlocuit spaniola cu chineză în exemplul meu de mai sus:

Dar IA m-aajunsdin urma, după o mare ruptură din 2016, cu aproape nici o limbă, aş putea traduce sistemul IA decât dezvoltat de echipa Google Brain.

Vă invit şi pe voi să încercaţi la https://translate.google.com, acum când citiţi cartea, şi să vedeţi dacă IA a lui Google s-a îmbunătăţit! Se prea poate să se fi întâmplat asta, căci există abordări promiţătoare

în mariajul reţelelor neuronale recurente profunde cu GOFAI pentru a crea o IA de procesare a limbajului care include un model al lumii.

Şanse şi provocări

Aceste trei exemple au fost doar o mostră, fiindcă IA progresează rapid pe mai multe fronturi importante. Apoi, deşi am menţionat doar două companii în aceste exemple, grupuri de cercetare din universităţi şi companii rivale nu se află mult în urmă. E mare agitaţie în catedrele de informatică, odată ce Apple, Baidu, DeepMind, Facebook, Google, Microsoft şi alte companii fac oferte tentante studenţilor, absolvenţilor şi profesorilor.

Exemplele pe care le-am dat n-ar trebui să ne facă să credem că istoria IA constă în perioade de stagnare întrerupte de realizări accidentale. Impresia mea este că asistăm de multă vreme la un progres relativ constant, iar mediile de informare anunţă câte o realizare atunci când e trecut pragul care permite o aplicaţie ce stârneşte imaginaţia sau un produs util. Cred deci că progresul susţinut al IA va continua încă mulţi ani. în plus, după cum am văzut în capitolul precedent, nu există vreun motiv fundamental pentru care progresul să nu continue până când IA va egala capacităţile umane la majoritatea capitolelor.

Se pune prin urmare întrebarea; ce efect va avea asta asupra noastră? Cum va modifica progresul IA pe termen scurt condiţia umană? Devine tot mai greu să susţii că IA este complet lipsită de obiective, anvergură, intuiţie, creativitate sau limbaj – trăsături considerate definitorii pentru om. Ca atare, chiar şi pe termen scurt, cu mult înainte ca IAG să ne egaleze la toate capitolele, IA ar putea avea un impact puternic asupra felului în care ne privim pe noi înşine, asupra a ceea ce putem face ajutaţi de IA şi asupra profesiunilor care ne rămân disponibile în competiţia cu IA. Va fi pozitiv sau negativ acest impact?

Ce şanse şi provocări pe termen scurt vor apărea?

Tot ce preţuim la civilizaţie este produsul inteligenţei umane, deci dacă o putem spori prin inteligenţa artificială, ne putem îmbunătăţi viaţa. Chiar şi progrese modeste ale IA ar aduce cu sine ameliorări în ştiinţă şi tehnologie, precum şi diminuări corespunzătoare ale accidentelor, maladiilor, injustiţiei, războaielor, corvezilor şi sărăciei. Dar pentru a ne bucura de aceste beneficii ale IA, fără a crea noi probleme, trebuie să răspundem la multe întrebări importante. De exemplu:

1. Cum putem face ca sistemele IA din viitor să fie mai robuste decât cele de azi, aşa încât să îndeplinească ce ne dorim fără să cedeze, să funcţioneze greşit sau să cadă pradă hackerilor?
2. Cum putem actualiza sistemele noastre juridice pentru a fi mai drepte şi mai eficiente, şi pentru a ţine ritmul cu peisajul digital aflat în schimbare rapidă?
3. Cum putem crea arme mai deştepte şi mai puţin dispuse să ucidă civili nevinovaţi, fără a declanşa o cursă a înarmărilor scăpată de sub control, cu arme autonome letale?
4. Cum ne putem spori prosperitatea prin automatizare, fără a-i lăsa pe oameni lipsiţi de venituri şi de un rost în viaţă? în restul capitolului de faţă vom cerceta pe rând fiecare dintre aceste întrebări. Ele îi vizează în principal pe informaticieni, jurişti, strategi militari şi economişti. Pentru a obţine însă răspunsurile de care avem nevoie în timp util, este nevoie ca toată lumea să participe la dezbatere, fiindcă, după cum vom vedea, provocările depăşesc toate frontierele tradiţionale – atât frontierele dintre profesii, cât şi cele dintre naţiuni.

Tehnologia informaţiei a avut deja un mare impact pozitiv practic asupra fiecărei activităţi umane – de la ştiinţă la finanţe, producţie, transport, ocrotirea sănătăţii, energie şi comunicaţii –, iar acest impact păleşte în comparaţie cu progresul pe care l-ar putea aduce IA. Dar cu cât ajungem să ne bazăm mai mult pe tehnologie, cu atât mai important devine ca ea să fie robustă şi demnă de încredere, făcând ceea ce vrem noi să facă.

în decursul istoriei umane, am învăţat din greşeli pentru a ne menţine benefică tehnologia. Am inventat focul, care ne-a dat adesea bătăi de cap, apoi am inventat extinctorul, ieşirea de incendiu, alarma de incendiu şi serviciul de pompieri. Am inventat automobilul, care a produs adesea accidente, apoi am inventat centurile de siguranţă, air bag-urile şi automobilele care se conduc singure. Până acum, tehnologiile noastre au provocat în genere accidente suficient de puţine şi de limitate pentru ca neajunsurile să fie eclipsate de avantaje. Cum însă elaborăm inevitabil tehnologii şi mai puternice, vom ajunge în punctul în care chiar şi un singur accident poate fi atât de catastrofal încât să eclipseze toate avantajele. Unii dau drept exemplu războiul nuclear global accidental, alţii pandemia produsă de bioinginerie, iar în capitolul următor vom vorbi despre temerea ca IA să conducă la dispariţia omenirii. Nu-i nevoie totuşi de asemenea exemple extreme pentru a ajunge la o concluzie esenţială: pe măsură ce tehnologia devine tot mai puternică, trebuie să ne bazăm tot mai puţin pe strategia de a învăţa din greşeli. Cu alte cuvinte, *trebuie să fim mai curând proactivi decât reactivi,* investind în cercetările legate de siguranţă pentru a evita ca accidentele să se producă fie şi o singură dată. De aceea societatea investeşte mai mult în siguranţa reactoarelor nucleare decât în siguranţa curselor de şoareci.

Acesta e şi motivul pentru care, după cum am văzut în capitolul i, la conferinţa din Puerto Rico s-a manifestat un puternic interes colectiv pentru cercetările în siguranţa IA. Computerele şi sistemele IA au cedat mereu, dar de data asta lucrurile stau altfel: IA pătrunde treptat în lumea reală, şi nu e o simplă neplăcere dacă provoacă prăbuşirea reţelei electrice, a bursei sau a sistemului armamentului nuclear. în restul acestei secţiuni voi vorbi despre cele patru mari domenii ale cercetărilor tehnice în siguranţa IA care domină dezbaterile actuale şi sunt studiate pretutindeni în lume: *verificare, validare, securitate* şi *control*[###] Pentru a evita să devenim plicticoşi şi arizi, vă propun să examinăm succesele şi eşecurile tehnologiei informaţiei din trecut în diferite sectoare, precum şi învăţămintele pe care le putem trage din ele şi provocările pe care le prezintă.

Deşi majoritatea acestor poveşti sunt vechi, implicând sisteme computerizate de mult depăşite, pe care aproape nimeni nu le-ar caracteriza drept IA, şi au provocat pagube neînsemnate, vom vedea că putem trage din ele învăţăminte preţioase pentru proiectarea unor viitoare sisteme IA sigure şi puternice, ale căror eşecuri s-ar putea dovedi cu adevărat catastrofale.

IA pentru explorarea spaţiului

Să începem cu un subiect care mi-e drag: explorarea spaţiului. Tehnologia calculatorului ne-a permis să trimitem oameni pe Lună şi nave fără echipaj uman să exploreze toate planetele sistemului solar, ba chiar şi să coboare pe suprafaţa lui Titan, satelitul

---

Dacă vreţi o hartă mai amănunţită a peisajului cercetărilorîn siguranţa IA, există una interactivă, elaborată printr-un efort comun dirijat de Richard Mallah de la FLI: https://futureoflife.org/landscape. *(N. a.)*

lui Saturn, şi pe o cometă. După cum vom vedea în capitolul 6, în viitor IA ar putea ajuta la explorarea altor sisteme solare şi galaxii – cu condiţia să fie lipsită de erori. Pe 4 iunie 1996, oamenii de ştiinţă care sperau să cerceteze magnetosfera Pământului sărbătoreau lansarea rachetei Ariane 5 a Agenţiei Spaţiale Europene, având Ia bord instrumentele construite de ei. Peste 37 de secunde însă Ie-a pierit zâmbetul, căci racheta a explodat într-un foc de artificii care costase sute de milioane de dolari.[8] Cauza era un software defect, care opera cu un număr prea mare pentru a intra în cei 16 biţi alocaţi lui.[9] Doi ani mai târziu, sonda spaţială Marş Climate Orbiter, lansată de NASA, a intrat din greşeală în atmosfera Planetei Roşii şi s-a dezintegrat fiindcă două părţi ale soft- ware-ului foloseau unităţi diferite pentru forţă, provocând o eroare de 445% în sistemul de control al motorului.[10] Era a doua eroare scump plătită de NASA: misiunea Mariner 1 către planeta Venus a explodat după lansarea de la Cape Cañaveral pe 22 iulie 1962, după ce software-ul pentru controlul zborului a fost dat peste cap de un semn de punctuaţie incorect.[11] Parcă pentru a dovedi că nu numai occidentalii stăpâneau arta de a lansa erori în spaţiu, misiunea sovietică Phobos 1 a eşuat pe 2 septembrie 1988. Era cea mai grea navă interplanetară lansată vreodată, cu obiectivul ambiţios de a depune un modul pe Phobos, satelitul lui Marte – dar absenţa unei cratime a provocat transmiterea comenzii „sfârşit-de-misiune“ către nava aflată în drum spre Marte, oprindu-i toate sistemele.[12]

Ce aflăm din aceste exemple este cât de important e ceea ce informaticienii numesc *verificare.* să te asiguri că software-ul satisface pe deplin toate cerinţele. Cu cât sunt în joc mai multe vieţi şi resurse, cu atât avem nevoie de o încredere mai mare că software-ul va funcţiona aşa cum dorim. Din fericire, IA poate ajuta Ia automatizare şi poate îmbunătăţi procesul de verificare. De exemplu, un

nucleu al sistemului de operare universal numit *seLţ* a fost recent verificat matematic, oferind o garanţie puternică împotriva defectării şi a operaţiilor periculoase: deşi nu a fost anunţat cu surle şi trâmbiţe ca Microsoft Windows şi Mac OS, poţi fi sigur că nu vei primi mesaje de eroare precum „ecranul albastru al morţii“ *(„the bine screen ofdeath“).* Agenţia pentru Proiecte de Cercetare Avansată pentru Apărare (Defence Advanced Research Projects Agency – DARPA) din Statele Unite a sponsorizat dezvoltarea unui set de instrumente de siguranţă cu sursă deschisă numit HACMS *(high-assurance cyber military sistems* – sisteme militare cibernetice de mare siguranţă) care s-a dovedit a fi sigur. O provocare importantă este de a face ca aceste instrumente să devină suficient de puternice şi uşor de folosit, aşa încât să fie larg răspândite. O altă provocare ţine de faptul că însăşi sarcina verificării devine mai dificilă când software-ul pătrunde în roboţi şi în noi medii, iar software-ul preprogramat tradiţional este înlocuit de sisteme IA care învaţă continuu, deci îşi schimbă comportamentul, după cum am văzut în capitolul 2.

IA pentru finanţe

Finanţele sunt alt domeniu care a fost transformat de tehnologia informaţiei, permiţând realocarea eficientă a resurselor oriunde în lume cu viteza luminii, şi înlesnind finanţarea avantajoasă pentru orice, de la ipoteci la companii startup. Progresul IA e probabil să ofere în viitor mari posibilităţi de profit din tranzacţii financiare: majoritatea deciziilor de cumpărare/vindere a acţiunilor sunt acum făcute automat de calculatoare, iar absolvenţii mei de la MIT sunt tentaţi cu salarii astronomice pentru a îmbunătăţi tranzacţionarea algoritmică.

Verificarea e importantă şi pentru software-ul financiar, ceea ce a aflat pe pielea ei firma americană

Knight Capital pe 1 august 2012, pierzând 440 de milioane de dolari in 45 de minute, după ce a instalat un software pentru tranzacţii neverificat.[13] „Crahul fulger“ de un bilion de dolari din 6 mai zoro merită menţionat din alt motiv. Deşi a provocat haos vreme de o jumătate de oră, până când pieţele s-au stabilizat, preţul acţiunilor unor companii mari precum Procter & Gambie fluctuând între un penny şi zoo 000 de dolari,[14] problema n-a fost declanşată de erori sau disfuncţii ale calculatoarelor pe care verificarea le-ar fi putut evita. Ea a avut în schimb drept cauză încălcarea aşteptărilor: programele de tranzacţionare automată de la numeroase companii au ajuns să opereze într-o situaţie neprevăzută, în care presupunerile lor nu mai erau valabile – de pildă, presupunerea că, dacă un computer al unei burse anunţa că o acţiune are preţul de un cent, atunci acţiunea valora într-adevăr un cent

Crahul fulger ilustrează importanţa a ceea ce informaticienii numesc *validare,* dacă verificarea pune întrebarea „Am construit oare corect sistemul?“, validarea pune întrebarea „Am construit oare sistemul corect?“[§§§] De exemplu: nu cumva sistemul se bazează pe presupuneri care s-ar putea să nu fie întotdeauna valabile? Dacă da, atunci cum poate fi perfecţionat pentru a trata mai bine incertitudinea?

IA pentru industrie

Evident, IA deţine un potenţial imens pentru îmbunătăţirea producţiei, controlând roboţi care sporesc deopotrivă eficienţa şi precizia. Imprimantele 3-D pot face acum prototipuri pentru orice, de la blocuri la dispozitive micromecanice mai mici decât un grăunte de sare.[15] în

§§§ Mai exact, verificarea întreabă dacă sistemul îndeplineşte specificaţiile, iar validarea dacă au fost alese specificaţiile corecte. *(N. a.)*

vreme ce roboţi industriali uriaşi construiesc automobile şi avioane, freze şi strunguri controlate de calculator, care pot fi cumpărate la preţuri accesibile, utilează nu doar fabricile, ci şi cele peste o mie de mici ateliere-laborator ale comunităţilor locale din lumea întreagă, unde entuziaştii îşi pot materializa ideile.[16] Cu cât avem însă mai mulţi roboţi în jurul nostru, cu atât e mai important să verificăm şi să validăm software-ul lor. Primul om ucis de un robot a fost Robert Williams, muncitor la uzinele Ford din Flat Rock, Michigan. În 1979, un robot care trebuia să coboare piese dintr-un spaţiu de depozitare s-a blocat, iar Williams s-a urcat să ia singur piesele. Robotul a început să funcţioneze pe tăcute şi i-a zdrobit capul, continuând să lucreze o jumătate de oră, până când ceilalţi muncitori au descoperit ce se întâmplase.[17] Următoarea victimă a roboţilor a fost Kenji Urada, inginer de întreţinere la uzinele Kawasaki din Akashi, Japonia. În 1981, în timp ce lucra la un robot defect, i-a atins din greşeală comutatorul şi a fost zdrobit de braţul hidraulic al robotului.[18] În 2015, un muncitor de la uzinele Volkswagen din Baunatal, Germania, punea în funcţiune un robot care trebuia să apuce piesele auto şi să le manevreze. Ceva s-a defectat, robotul l-a apucat şi l-a strivit de un banc metalic.[19]

Deşi aceste accidente sunt tragice, e important de observat că ele reprezintă o fracţiune minusculă din ansamblul accidentelor industriale. De fapt, accidentele industriale nu s-au înmulţit, ci s-au *rărit* odată cu perfecţionarea tehnologiei, scăzând în Statele Unite de la circa 14 000 de decese în 1970 la 4 821 în 2014.[20] Cele trei accidente pomenite aici arată că dotarea unor maşini comune cu inteligenţă ar trebui să sporească siguranţa industrială, învăţându-i pe roboţi să fie mai atenţi în apropierea oamenilor. Toate cele trei accidente puteau fi evitate printr-o validare mai bună: roboţii au produs vătămări nu din rea voinţă sau din pricina vreunei erori

din program, ci fiindcă au făcut presupuneri greşite – că persoana nu se afla acolo sau că era o piesă auto.

Figura 3.3 Dacă roboţii industriali sunt scumpi şi dificil de programat, există acum o tendinţă către roboţi mai ieftini, echipaţi cu IA, care pot învăţa ce să facă de la muncitori fără experienţă în programare.

Dacă IA poate salva multe vieţi în industrie, ea poate salva încă şi mai multe vieţi în transport. Numai accidentele de maşină au ucis peste 1,2 milioane de oameni în 2015, iar accidentele aviatice, feroviare şi navale au ucis împreuna alte câteva mii. în Statele Unite, unde sunt standarde înalte de siguranţă, accidentele vehiculelor motorizate au făcut în 2016 circa 35 000 de victime, de şapte ori mai mult decât toate accidentele industriale la un loc.[ZI] într-o discuţie pe această temă la întrunirea anuală a Asociaţiei pentru Progresul Inteligenţei Artificiale din 2016, desfăşurată la Austin, Texas, informaticianul israelian Moshe Vardi a spus că IA nu doar poate reduce accidentele auto, dar *trebuiesc* facă: „Este un imperativ moral!“ a exclamat el. Cum majoritatea accidentelor de maşină sunt provocate de erori umane, se crede că automobilele echipate cu IA care se conduc singure pot elimina cel puţin 90% din decesele rutiere, iar acest optimism încurajează apariţia efectivă a unor asemenea maşini pe drumuri. Elon Musk estimează că viitoarele automobile care se conduc singure vor fi nu doar mai sigure, dar vor aduce şi bani proprietarilor lor atunci când aceştia n-au nevoie de ele, putând concura cu Uber şi Lyft.

Până acum, automobilele care se conduc singure s-au dovedit mai sigure decât şoferii umani, iar accidentele petrecute subliniază importanţa şi dificultatea validării. Prima ciocnire minoră provocată de un asemenea automobil Google a avut loc pe 14 februarie 2014, din cauza unei presupuneri incorecte privind un autobuz: că şoferul va ceda atunci când maşina a apărut în faţa lui. Primul accident mortal provocat de un automobil Tesla, care a intrat în remorca unui camion pe 7 mai 2016, a fost provocat de două presupuneri greşite?[2] că latura albă strălucitoare a remorcii era o parte a cerului strălucitor şi că şoferul (care se pare că urmărea un film cu Harry

Potter) era atent şi ar fi intervenit în caz de urgenţă.[****]

Uneori însă, buna verificare şi validarea nu sunt de ajuns pentru a evita accidentele, fiindcă avem nevoie şi de un bun *control:* capacitatea operatorului uman de a urmări sistemul şi de a-i schimba comportamentul dacă e necesar. Pentru ca asemenea sisteme cu „om în buclă“ să funcţioneze bine e esenţială eficacitatea comunicării om-maşină. în această idee, o lumină roşie pe tabloul de bord te va alerta dacă din greşeală ai lăsat deschis portbagajul maşinii. Când feribotul britanic pentru automobile Herald of Free Enterprise a părăsit portul Zeebrugge, pe 6 martie 1987, cu rampa de la provă coborâtă, n-a existat vreo lumină de alarmă sau alt semnal de avertizare pentru căpitan, iar feribotul s-a răsturnat în scurt timp, făcând 193 de victime?[3]

Altă greşeală tragică de control care putea fi evitată printr-o mai bună comunicare om-maşină s-a petrecut în noaptea de 1 iunie 2009, când zborul Air France 447 s-a prăbuşit în Oceanul Atlantic, toţi cei 228 de oameni aflaţi la bord murind. Conform raportului oficial, „echipajul n-a înţeles că pierdeau din portanţă, şi deci n-a aplicat manevra de redresare“ – coborârea botului avionului – decât prea târziu. Experţii în securitatea aviatică au afirmat că accidentul putea fi evitat dacă în cabină ar fi existat un indicator al „unghiului de atac“ care să le arate piloţilor că botul era îndreptat prea mult în sus.[24]

Când zborul Air France 248 s-a prăbuşit în Munţii Vosgi, în apropiere de Strasbourg, pe 20 ianuarie 1992,87 de persoane pierzân- du-şi viaţa, cauza n-a fost lipsa de comunicare om-maşină, ci o interfaţă derutantă. Piloţii au introdus „33“ pe tastatură, pentru că voiau să coboare la un unghi de 3,3 grade, dar pilotul automat a interpretat că asta reprezintă 3 300 de picioare pe minut, fiindcă se

**** Chiar luând în calcul acest accident, introducerea autopilotului Tesla a redus cu 40% numărul accidentelor: http://tinyurl.com/teslasafety.

afla într-un mod diferit, iar ecranul era prea mic pentru a afişa modul şi a permite piloţilor să-şi dea seama de greşeală.

IA pentru energie

Tehnologia informaţiei a făcut minimi pentru generarea şi distribuţia energiei, cu algoritmi sofisticaţi care echilibrează producţia şi consumul în reţelele electrice, şi cu sisteme de control sofisticate care menţin funcţionarea sigură şi eficientă a centralelor
electrice. Viitoarele progrese ale IA vor face probabil şi mai inteligentă „reţeaua inteligentă“, pentru a se adapta optim cererii şi ofertei schimbătoare până la nivelul individual al panourilor solare de pe acoperişuri şi al sistemelor casnice de stocare a energiei. Totuşi, pe 14 august 2003, o pană de curent a afectat circa 55 de milioane de americani şi canadieni, mulţi dintre ei rămânând fără energie câteva zile. Şi aici, cauza principală a fost comunicarea defectuoasă om-maşină: o eroare în software a împiedicat sistemul de alarmă dintr-o cameră de comandă din Ohio să alerteze operatorii că e nevoie să redistribuie energia înainte ca o problemă minoră (linii de transmisie supraîncărcate prinse într-un frunziş des) să ia amploare şi să scape de sub control?[5]

Topirea parţială a reactorului nuclear din centrala Three Mile Island, Pennsylvania, din 28 martie 1979, a adus pierderi de un miliard de dolari şi a declanşat o reacţie violentă împotriva energiei nucleare. Raportul final asupra accidentului a identificat mai multe cauze, între care confuzia provocată de slaba comunicare om-maşină?[6] în particular, becul de avertizare despre care operatorii credeau că arată dacă o valvă de siguranţă e închisă sau deschisă, indica doar dacă a fost transmis un semnal pentru a închide valva – aşa încât operatorii nu şi-au dat seama că valva rămăsese înţepenită deschisă.

Aceste accidente legate de energie şi transmiterea ei ne arată că, pe măsură ce punem IA să controleze şi mai multe sisteme fizice, trebuie să facem eforturi de cercetare serioase vizând nu doar funcţionarea corectă a maşinilor pe cont propriu, dar şi colaborarea eficientă cu controlorii lor umani. Progresele IA vor aduce cu sine atât crearea unor interfeţe perfecţionate pentru utilizator în vederea partajării informaţiei, cât şi soluţii pentru alocarea optimă a sarcinilor în echipele om-calculator – de exemplu, identificarea situaţiilor în care controlul trebuie transferat, pentru aplicarea eficientă a judecăţii umane în deciziile cruciale, în loc să distragi controlorii umani cu un flux de informaţii neimportante.

IA pentru ocrotirea sănătăţii

IA are un potenţial imens pentru îmbunătăţirea asistenţei medicale. Digitalizarea fişelor medicale a permis deja doctorilor şi pacienţilor să ia hotărâri mai bune şi mai rapide, şi să primească instantaneu ajutorul experţilor de oriunde în lume la diagnosticarea imaginilor digitale. Cei mai buni experţi în acest tip de diagnostic ar putea deveni curând sistemele IA, dat fiind progresul rapid în imagistica computerizată şi învăţarea profundă. De pildă, un studiu olandez din 2015 arăta că diagnosticarea computerizată a cancerului de prostată folosind imagistica prin rezonanţă magnetică (IRM) era la fel de bună ca cel al radiologilor umani,[27] iar un studiu Stanford din 2016 arăta că IA poate diagnostica mai bine decât anatomopatologii umani cancerul pulmonar, folosind imaginile de la microscop?[8] Dacă învăţarea automată ar putea ajuta la descoperirea relaţiilor dintre gene, boli şi răspunsul la tratament, ar revoluţiona actul medical, ar îmbunătăţi sănătatea animalelor din ferme şi ar duce la obţinerea de recolte mai rezistente. Apoi, roboţii ar putea deveni chirurgi mai precişi şi mai de încredere decât oamenii, chiar şi fără a folosi IA avansată. O mare

diversitate de intervenţii chirurgicale robotice au fost efectuate cu succes în ultimii ani, precizia, miniaturizarea şi inciziile mai mici făcând să scadă pierderile de sânge, durerile şi perioada de recuperare.

Din nefericire însă, câteva cazuri tragice ne atrag atenţia asu pra importanţei unui software robust şi în domeniul sănătăţii. De exemplu, instalaţia canadiană de radioterapie Therac-25 a fost proiectată pentru a trata pacienţii cu cancer în două moduri: fie cu un fascicul de electroni de putere scăzută, fie cu un fascicul de raze X de mare putere, un megavolt, concentrat doar asupra ţintei cu ajutorul unui ecran protector special. O eroare dintr-un software neverificat i-a făcut pe tehnicieni să trimită fasciculul de un megavolt, când ei credeau că administrează fasciculul de putere scăzută, şi fără ecranul protector, provocând decesul câtorva pa- cienţi?[9] încă şi mai mulţi pacienţi au murit din cauza supradozei de radiaţii la Institutul Oncologic Naţional din Panama, unde echipamentul de radioterapie cu cobalt-6o a fost programat cu perioade de expunere prea lungi în 2000 şi 2001, din cauza unei interfeţe care nu fusese corect validata.[30] Conform unui raport recent,[31] în Statele Unite, între 2000 şi 2013, accidentele provocate de chirurgia robotizată au dus la 144 de decese şi 1391 de vătămări, problemele fiind legate nu numai de hardware – arcuri electrice şi arderi, sau piese desprinse din instrumente căzând în pacient –, dar şi de software – mişcări necontrolate şi deconectări spontane.

Vestea bună este că restul celor aproape două milioane de intervenţii chirurgicale robotizate incluse în raport s-au desfăşurat perfect, iar roboţii par să fie destul de siguri în chirurgie. Conform unui studiu guvernamental american, malpraxisul medical în spitale are drept efect peste roo 000 de decese anual numai în Statele Unite,[32] prin urmare imperativul moral de a dezvolta IA pentru medicină e probabil încă mai puternic decât pentru auto-

mobile care se conduc singure.

IA pentru comunicaţii

Industria comunicaţiilor este pesemne cea în care calculatoarele au avut până acum cel mai mare impact După introducerea tablourilor de comandă telefonice computerizate în anii ’50, a inteme- tului în anii ’60 şi a World Wide Web în 1989, miliarde de oameni comunică, fac cumpărături, citesc ştiri, se uită la filme sau joacă jocuri online, obişnuiţi să aibă informaţia din întreaga lume la un clic distanţă – şi adesea gratuit. Acel *internet al lucrurilor (internet of things)* aflat în plin avânt promite să crească eficienţa, precizia, comoditatea şi beneficiul economic aducând online totul, de la lămpi, termostate şi frigidere la transpondere cu biocipuri şi animale de fermă.

Aceste succese spectaculoase în conectarea lumii i-au adus pe informaticieni în faţa unei a patra provocări: trebuie să îmbunătăţească nu numai verificarea, validarea şi controlul, dar şi *securitatea* împotriva software-ului malefic (malware) şi a hackerilor. Dacă problemele menţionate mai sus erau rezultatul unor greşeli neintenţionate, securitatea vizează acţiuni *deliberate.* Primul malware care a atras atenţia mediilor de informare a fost aşa-nu- mitul „vierme informatic Morris“, lansat pe 2 noiembrie 1988, care exploata erorile din sistemul de operare UNIX. A fost, chipurile, o încercare greşit orientată de a număra câte calculatoare erau online, şi deşi a infectat şi blocat circa 10% din cele 60 000 de calculatoare care compuneau pe atunci internetül, asta nu l-a împiedicat pe creatorul său, Róbert Morris, să ajungă în cele din urmă profesor de informatică la MIT.

Alte malware-uri exploatează vulnerabilităţile nu doar ale software ului, ci şi ale oamenilor. Pe 5 mai 2000, parcă pentru a sărbători ziua mea de naştere, lumea a primit e

mailuri cu subiectul „ILOVEYOU“ (TEIUBESC) de la cunoştinţe şi colegi, iar acei utilizatori ai Microsoft Windows care au deschis documentul ataşat „LOVE-LETTER-FOR-YOU.tXt.vbs“ (SCRISOARE DE-DRAGOSTE-PEN- TRU-TINE) au lansat fără să vrea un program care le-a avariat calculatorul şi a retrimis e mailul tuturor din lista lor de adrese. Creat de doi tineri programatori din Filipine, acest vierme informatic a infectat aproximativ 10% din internet, la fel cum o făcuse viermele lui Morris, dar fiindcă internetül era deja mult mai vast, a devenit una dintre cele mai grave infectări din toate timpurile, afectând peste 50 de milioane de calculatoare şi provocând o pagubă de peste 5 miliarde de dolari. După cum din nefericire ştiţi, internetül e în continuare infestat cu nenumărate tipuri de malware, pe care experţii în securitate îl clasifică în viermi, troieni, viruşi şi alte categorii cu rezonanţe înspăimântătoare, iar daunele provocate se întind pe un spectru larg, de la afişarea unor mesaje inocente la ştergerea fişierelor, furtul informaţiilor personale, spionare şi deturnarea calculatorului pentru a trimite sp amuri.

Dacă malware-ul vizează orice calculator găseşte, *hackerii* atacă anumite ţinte de interes – între care, recent, Target, TJ Maxx, Sony Pictures, Ashley Madison, compania petrolieră saudită Aramco şi Comitetul Naţional Democrat al Statelor Unite. Jaful însă pare să fie şi mai spectaculos. în 2008, hackerii au furat 130 de milioane de numere ale cărţilor de credit şi alte informaţii despre cont de la sistemele de plată Heartland, iar în 2013 au pătruns în peste un miliard (!) de conturi de e-mail Yahoo!.[33] în 2014, de la Oficiul Guvernamental de Management al Personalului din Statele Unite s-au sustras informaţii din dosarele personale şi cererile de angajare a peste 21 de milioane de oameni, între care angajaţi cu acces la informaţii confidenţiale, şi au fost obţinute amprentele unor agenţi sub acoperire.

Mă crucesc de fiecare dată când citesc despre un

sistem prezentat a fi 100% sigur şi inatacabil. Totuşi, viitoarele sisteme IA trebuie să fie „inatacabile“ înainte să preia comanda infrastructurii critice sau a armamentului, de pildă – aşa încât creşterea rolului IA în societate ridică miza securităţii calculatoarelor. Dacă unii hackeri exploatează naivitatea umană sau vulnerabilităţile complexe din software-ul recent lansat, alţii se conectează neautorizat la calculatoare străine speculând erori simple de programare care au rămas mult timp neobservate. Bug-ul „Heartbleed“ a supravieţuit din 2012 până în 2014 într-una dintre cele mai populare biblioteci de software pentru comunicaţii sigure între calculatoare, iar bug-ul „Bashdoor“ s-a aflat chiar în sistemul de operare al calculatoarelor Unix din 1989 până în 2014. Asta înseamnă că instrumentele IA pentru ameliorarea verificării şi validării vor spori şi securitatea.

Din păcate, sisteme IA perfecţionate pot fi folosite şi pentru a găsi noi vulnerabilităţi şi a-i înzestra pe hackeri cu arme şi mai sofisticate. închipuie-ţi, de exemplu, că primeşti un e-mail-cap- cană neobişnuit de personalizat care încearcă să te convingă să divulgi informaţii personale. E trimis de pe contul prietenei tale de către IA a unui hacker şi îi imită stilul pe baza analizei altor mesaje trimise de ea, incluzând o mulţime de informaţii personale despre tine din alte surse. Te-ai lăsa păcălit? Dar dacă mesa- jul-capcană ar părea să provină de la compania care ţi-a eliberat cârdul de credit şi e urmat de un apel telefonic în care îţi vorbeşte o voce umană prietenoasă despre care nu ştii că e generată prin IA? în cursa înarmărilor dintre atac şi apărare din securitatea informatică nu prea avem indicii că apărarea ar câştiga.

LEGI

Noi, oamenii, suntem animale sociale care am subjugat celelalte specii şi am cucerit Pământul graţie capacităţii noastre de a coopera. Am conceput legi pentru a stimula

şi înlesni cooperarea, aşa încât, dacă IA poate îmbunătăţi sistemele noastre juridice şi de guvernare, ne-ar ajuta să cooperăm încă şi mai eficient decât până acum şi să ne atingem potenţialul maxim. Şi există o mulţime de lucruri de îmbunătăţit, atât în aplicarea, cât şi în elaborarea legilor, aşa încât vă propun să Ie examinăm pe ambele.

Ce îţi vine în minte când te gândeşti la sistemul juridic din ţara ta? Termene lungi, costuri ridicate şi, uneori, nedreptate – şi nu eşti singurul. N-ar fi oare minunat dacă primele cuvinte care ţi-ar veni în minte ar fi „eficienţă" şi „corectitudine"? Din moment ce procesul juridic poate fi privit în mod abstract ca un calcul, introducând informaţii despre dovezi şi legi şi extragând o decizie, unii visează să-l automatizeze complet cu *robojudecători:* sisteme IA care aplică neobosit aceleaşi standarde juridice înalte la orice hotărâre judecătorească, fără a cădea pradă erorilor umane cum sunt părtinirea, oboseala şi ignorarea celor mai recente cunoştinţe tehnice.

Robojudecători

Byron De La Beckwith jr. a fost condamnat în 1994 pentru asasinarea, în 1963, a militantului pentru drepturile civile Medgar Evers, dar două jurii diferite din Mississippi, compuse numai din albi, îl achitaseră la un an după crimă, deşi dovezile fizice erau în esenţă aceleaşi.[34] Din nefericire, istoria justiţiei e plină de asemenea verdicte influenţate de culoarea pielii, gen, orientare sexuală, religie, naţionalitate şi alţi factori. în principiu, robojudecătorii ar putea face ca pentru prima dată în istorie toţi oamenii să fie cu adevărat egali în faţa legii: ei ar fi programaţi să fie identici şi să trateze pe toată lumea în mod egal, aplicând transparent legea, fără părtinire.

Robojudecătorii ar putea elimina şi preferinţele umane

accidentale, neintenţionate. De pildă un controversat studiu asupra judecătorilor israelieni afirma că aceştia dau verdicte semnificativ mai aspre atunci când le este foame: dacă imediat după micul dejun refuză 35% din cererile de eliberare condiţionată, chiar înainte de prânz ei refuză 85%.[35] Alt neajuns al judecătorilor umani este că s-ar putea să nu aibă timp să studieze toate detaliile unui proces. Robojudecătorii pot fi însă cu uşurinţă copiaţi, fiindcă în esenţă ei nu sunt decât un software, aşa încât toate procesele aflate în aşteptare pot fi judecate în paralel, fiecare având propriul lui robojudecător pentru atât timp cât e nevoie. în fine, dacă judecătorilor umani le e imposibil să stăpânească toate cunoştinţele tehnice cerute de fiecare caz în parte, de la dispute spinoase pentru brevete la misterele unor crime depinzând de cele mai avansate cunoştinţe de medicină legală, viitorii robojudecătorî vor avea memorie şi capacitate de învăţare practic nelimitate.

într-o bună zi, asemenea robojudecătorî ar putea fi deci mai eficienţi şi mai corecţi graţie nepărtinirii, competenţei şi transparenţei. Eficienţa i-ar face să devină încă şi mai corecţi: accelerând procesul şi îngreunând încercările avocaţilor versaţi de a influenţa verdictul, actul de justiţie ar deveni mult mai ieftin, ceea ce ar spori şansele ca un om sărac sau o companie startup să câştige în faţa unui miliardar sau a unei corporaţii multinaţionale pentru care lucrează o armată de avocaţi.

Dar ce se întâmplă dacă robojudecătorii au bug-uri sau sunt victimele hackerilor? Ambele primejdii au afectat deja maşinile de vot automat, iar atunci când miza e închisoarea sau o sumă enormă, riscurile unui atac cibernetic sunt şi mai mari. Chiar daca IA devine suficient de robustă pentru a ne da încredere că roboju- decătorul foloseşte algoritmul legal, va avea oare toată lumea senzaţia că înţelege îndeajuns logica raţionamentelor lui

pentru a-i respecta verdictul? Această problemă e accentuată de recentul succes al reţelelor neuronale, care adesea depăşesc cu mult algoritmii IA tradiţionali uşor de înţeles, preţul plătit fiind neinteligi- bilitatea. Dacă inculpaţii vor să ştie *de ce* au fost condamnaţi, nu ar avea oare ei dreptul la un răspuns mai bun decât „am antrenat sistemul cu o mulţime de date, iar asta a hotărât“? în plus, studii recente au arătat că dacă antrenezi un sistem neuronal de învăţare profundă cu cantităţi mari de date privind deţinuţii, el poate prezice cine e probabil să recidiveze (şi ar trebui deci să i se refuze eliberarea condiţionată) mai bine decât judecătorii umani. Ce se întâmplă însă dacă sistemul găseşte că recidivismul e corelat statistic cu sexul sau rasa deţinutului – ar fi considerat un robojude- cător sexist sau rasist care trebuie reprogramat? Un studiu din 2016 susţine că software-ul folosit în Statele Unite pentru predic- ţia recidivismului îi defavoriza pe afro-americani şi a contribuit la sentinţe nedrepte^[6] Acestea sunt aspecte importante care trebuie analizate şi discutate pentru a ne asigura că IA rămâne benefică. Nu suntem puşi în faţa unei hotărâri de tipul totul-sau-nimic în privinţa robojudecătorilor, ci ar trebui să hotărâm în ce măsură şi în ce ritm vrem să folosim IA în sistemul nostru juridic. Vrem ca judecătorii umani să fie ajutaţi de sisteme de decizie bazate pe IA, la fel ca medicii de mâine? Vrem să mergem mai departe şi să avem decizii ale robojudecătorilor, iar apelul să fie lăsat în seama judecătorilor umani, sau vrem să mergem până la capăt, dând maşinilor ultimul cuvânt chiar şi în cazul pedepsei cu moartea?

Controverse juridice

Am examinat până acum doar *aplicarea* legii, să ne întoarcem Ia *conţinutul* ei. Există un larg consens cu

privire la faptul că legile noastre trebuie să evolueze pentru a ţine ritmul cu tehnologia. De exemplu, cei doi programatori care au creat viermele ILOVEYOU, provocând pagube de miliarde de dolari, au fost achitaţi de toate acuzaţiile pentru că în Filipine nu existau pe atunci legi în privinţa malware-ului. De când ritmul progresului tehnologic s-a accelerat, legile trebuie actualizate şi mai rapid, căci au tendinţa să fie defazate. Angajarea tot mai multor informaticieni experţi în învăţământul juridic şi în guverne ar fi probabil o soluţie bună. Ar trebui însă oare să apară sisteme de luare a deciziilor bazate pe IA pentru votanţi şi legiuitori, urmate de robolegiuitori în toată puterea cuvântului?

Cum ar trebui modificate legile noastre pentru a reflecta progresul IA e o fascinantă temă de dezbatere. Un aspect se leagă de tensiunea dintre libertatea informaţiei şi caracterul ei privat. Partizanii libertăţii susţin că cu cât avem mai puţină confidenţialitate, cu atât vor avea tribunalele mai multe probe, iar verdictele lor vor fi mai corecte. De pildă, dacă guvernul foloseşte dispozitivele electronice ale oricui pentru a înregistra unde e, ce tastează la calculator, ce spune şi ce face, multe infracţiuni ar fi rapid lămurite, iar altele ar fi împiedicate. Apărătorii confidenţialităţii răspund că nu-şi doresc un stat orwellian care te supraveghează, şi, chiar dacă şi-ar dori, există riscul să se transforme într-o dictatură totalitară de proporţii gigantice. în plus, tehnicile de învăţare automată s-au perfecţionat în analizarea datelor cerebrale ale scanărilor IRMf pentru a determina ia ce se gândeşte un om şi, în particular, pentru a determina dacă spune sau nu advărtuU[7] Dacă tehnologia de scanare cerebrală asistată de IA ar deveni ceva obişnuit în tribunale, stabilirea faptelor într-un proces ar fi mult simplificată şi accelerată, iar verdictele ar fi mai corecte. Apărătorii confidenţialităţii ar putea fi însă îngrijoraţi că

asemenea sisteme ar comite din când în când erori, şi mai ales că guvernul ne-ar spiona nelimitat. Guvernele care nu susţin libertatea de gândire ar putea folosi această tehnologie pentru a-i incrimina pe cei care au anumite credinţe şi opinii. Unde ai trasa *tu* graniţele dintre justiţie şi confidenţialitate şi dintre protejarea societăţii şi protejarea libertăţii personale? Indiferent unde ai trasa-o, se va deplasa ea oare treptat, dar inexorabil, către reducerea confidenţialităţii pentru a compensa faptul că dovezile devin mai uşor de falsificat? De exemplu, odată ce IA poate genera imagini realiste cu tine comiţând crime, ai vota pentru un sistem prin care guvernul urmăreşte în permanenţă pe toată lumea şi care ţi-ar oferi un alibi imbatabil, dacă ai avea nevoie de el?

Altă controversă captivantă se leagă de reglementarea IA sau, mai general, de felul în care factorii politici ar trebui să stimuleze cercetările IA pentru a maximiza şansele unui rezultat benefic. Unii cercetători refuză toate formele de reglementare a dezvoltării IA, susţinând că acestea ar întârzia inutil inovaţiile necesare (de pildă, automobile care, conducându-se singure, ar salva vieţi) şi ar face ca cercetarea IA de vârf să devină subterană şi/sau ar transfera-o în alte ţări, cu guverne mai permisive. La conferinţa din Puerto Rico, Elon Musk a susţinut că din partea guvernelor avem nevoie de înţelegere, nu de supraveghere: experţi ocupând poziţii guvernamentale care să urmărească progresul IA şi să-l îndrume. El a mai spus că reglementările guvernamentale pot să stimuleze uneori progresul, nu să-l înăbuşe: de exemplu, dacă standardele de siguranţă guvernamentale pentru automobilele care se conduc singure pot duce Ia scăderea numărului accidentelor, atunci e mai puţin probabilă o reacţie negativă din partea publicului, iar adoptarea noii tehnologii poate fi accelerată. Cele mai precaute companii LA ar favoriza prin urmare

reglementările care îi obligă pe competitorii mai puţin scrupuloşi să se conformeze standardelor înalte de siguranţă.

O altă controversă juridică interesantă se referă la drepturile pe care le-am acorda maşinilor. Dacă automobilele care se conduc singure ar înjumătăţi cele 32 000 de accidente rutiere mortale anuale din Statele Unite, poate că producătorii auto n-ar primi 16 000 de scrisori de mulţumire, ci 16 000 de trimiteri în judecată. Aşadar, dacă un automobil care se conduce singur provoacă un accident, cine e răspunzător – pasagerii, proprietarul sau producătorul? Juristul David Vladeck propune un al patrulea răspuns: automobilul însuşi! Mai exact, ideea e ca automobilelor care se conduc singure să li se permită (şi să li se ceară) să aibă o asigurare auto. Astfel, modelele cu un „cazier" excelent ar beneficia de mari reduceri, ajungându-se la sume mai mici decât cele disponibile pentru şoferii umani, în vreme ce modelele de calitate slabă ar avea poliţe de asigurare care să le facă prohibitiv de scumpe.

Dar dacă maşinilor cum sunt automobilele li se permite să aibă poliţe de asigurare, ar trebui să aibă ele voie şi să deţină bani sau proprietăţi? Dacă da, nu există nici un argument juridic care să împiedice calculatoarele isteţe să câştige bani la bursă şi să-i folosească pentru a cumpăra servicii online. Odată ce un calculator începe să-i plătească pe oameni ca să lucreze pentru el, acesta poate realiza orice pot realiza oamenii Dacă sistemele IA reuşesc în cele din urmă să se priceapă mai bine decât oamenii la investiţii (ceea ce se întâmplă deja în anumite domenii), s-ar putea ajunge Ia situaţia în care cea mai mare parte a economiei noastre să fie deţinută şi controlată de maşini. Asta ne dorim? Pare o perspectivă îndepărtată, însă majoritatea economiei noastre este de pe-acum deţinută de o altă formă de entitate non-umană:

corporaţiile, care adesea sunt mai puternice decât orice persoană din cadrul lor şi pot într-o anumită măsură să-şi ia viaţa pe cont propriu.

Dacă acorzi maşinilor dreptul la proprietate, le-ai acorda şi dreptul de a vota? în caz că da, ar trebui oare ca fiecare program de calculator să aibă un vot, deşi pentru el e banal să-şi facă bilioane de copii ale lui însuşi în *cloud* (nor), dacă e suficient de bogat, hotărând astfel rezultatul tuturor alegerilor? în caz că nu, atunci pe ce bază morală defavorizăm minţile maşinilor în raport cu minţile umane? Are vreo importanţă dacă minţile maşinilor sunt sau nu conştiente, în sensul că posedă o experienţă subiectivă ca a noastră? Vom examina în detaliu problemele legate de controlul calculatorului asupra lumii noastre în capitolul următor, iar cele legate de conştiinţa maşinii în capitolul 8.

ARME

Din vremuri imemoriale omenirea a suferit din pricina foametei, a bolilor şi a războaielor. Am văzut deja cum poate IA reduce foametea şi bolile, dar ce putem spune despre războaie? Unii susţin că armele nucleare descurajează războaiele dintre ţările care le deţin, fiindcă sunt atât de înspăimântătoare. Ar trebui aşadar să permitem tuturor naţiunilor să construiască arme şi mai înspăimântătoare bazate pe IA, în speranţa că nu vor mai exista niciodată războaie? Dacă nu eşti convins de acest argument şi crezi că războaiele viitoare sunt inevitabile, ce părere ai de folosirea IA pentru a face ca aceste războaie să fie mai umane? Dacă războaiele se reduc la maşini luptând împotriva maşinilor, atunci nu mai sunt ucişi soldaţi sau civili umani. în plus, viitoarele drone acţionate de IA şi alte sisteme de arme autonome *(autonomous weapon systems* – AWS; numite de oponenţii

lor şi „roboţi ucigaşi“) este de sperat să fie mai corecte şi mai raţionale decât soldaţii umani: echipate cu senzori supraumani şi netemându-se că vor fi ucise, ar rămâne reci, calculate şi echilibrate chiar şi în toiul bătăliei, fiind mai puţin probabil să ucidă în mod accidental civili.

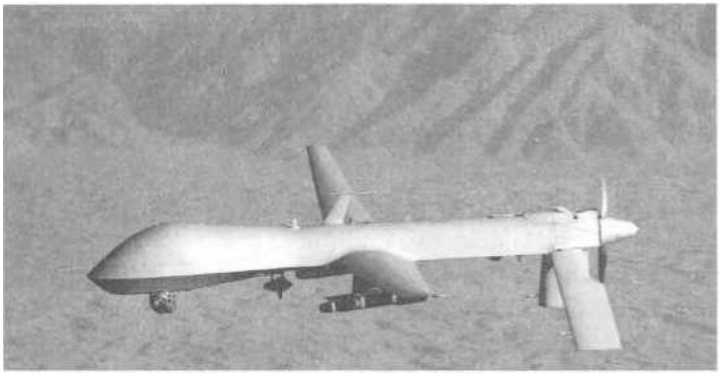


Figura 3.4 Dacă dronele militare de azi (cum e această MQ-i Predator a Forţelor Aeriene ale Statelor Unite) sunt controlate de la distanţă de oameni, viitoarele drone acţionate de IA ar putea lăsa oamenii pe dinafară, folosind un algoritm care să hotărască ce ţintă trebuie atinsă.

Un om în circuitul deciziei

Dar dacă sistemele automate conţin erori, sunt derutante sau nu se comportă cum ne aşteptăm? Sistemul american Phalanx pentru crucişătoarele din clasa Aegis detectează, urmăreşte şi atacă rachete sau avioane ameninţătoare. Vincennes era un crucişător american botezat Robocruiser datorită automatizărilor cu care era echipat, iar pe 3 iulie 1988, în toiul unei încăierări cu canoni- erele iraniene din timpul războiului dintre Iran şi Irak, radarul

lui a detectat un avion care se apropia. Căpitanul William Rodgers III a dedus că erau atacaţi în picaj de un avion de vânătoare iranian F-14 şi a dat sistemului Aegis aprobarea să deschidă focul. Nu şi-a dat seama însă că a doborât

zborul Iran Air 655, un avion iranian civil de pasageri, ucigându-i pe toţi cei 290 de oameni aflaţi la bordul lui şi provocând indignarea întregii lumi. Investigaţiile ulterioare au scos la iveală că de vină era interfaţa om-maşină, care nu arăta automat ce puncte de pe ecranul radarului erau avioane civile (zborul 655 îşi urma cursul zilnic regulat, transponderul lui de aeronavă civilă fiind în funcţiune) şi ce puncte coborau (ca pentru atac) sau urcau (aşa cum se întâmpla cu zborul 655, după decolarea din Teheran). Când sistemului automat i s-au cerut informaţii despre misteriosul avion, a raportat că „se află în coborâre“, fiindcă asta făcea un alt avion, căruia din greşeală îi fusese realocat un număr folosit de forţele navale pentru a urmări avioanele: ceea ce cobora era un avion american care patrula departe, în Golful Oman.

în acest exemplu, decizia finală a fost luată de un om aflat în circuit, care, sub presiunea timpului, a avut prea multă încredere în ce-i spunea sistemul automat. Conform responsabililor militari din lumea întreagă, toate sistemele de armament desfăşurate deocamdată includ un om în circuitul deciziei, cu excepţia capcanelor primitive de tipul minelor terestre. Sunt însă concepute acum arme cu adevărat autonome, care selectează şi atacă ţinte pe cont propriu. Din punct de vedere militar e tentant să scoţi toţi oamenii din circuitul deciziei pentru a spori viteza: într-o luptă dintre o dronă complet autonomă, care poate răspunde instantaneu, şi o dronă care reacţionează mai lent, fiindcă e controlată de oameni de la mare distanţă, care credeţi că ar învinge?

Pe de altă parte însă, au existat situaţii în care am fost extrem de norocoşi că a existat un om în circuitul deciziei. Pe 27 octombrie 1962, în timpul crizei rachetelor din Cuba, unsprezece distrugătoare şi portavionul Randolph încolţiseră un submarin sovietic B-59 în apropierea coastei cubaneze, în apele internaţionale din afara zonei

americane de „carantină“. Ceea ce nu se ştia era că temperatura la bordul lui ajunsese la 45°C, pentru că bateriile submarinului se descărcaseră, iar aerul condiţionat nu mai funcţiona. Intoxicaţi cu dioxid de carbon, mulţi membri ai echipajului leşinaseră. Trecuseră câteva zile de când nu mai comunicaseră cu Moscova şi nu ştiau dacă nu cumva al Treilea Război Mondial începuse deja. Americanii au lansat atunci mici grenade antisubmarin, şi au spus Moscovei că nu voiau decât să oblige submarinul să iasă la suprafaţă şi să plece. „Credeam că, gata, s-a sfârşit“, avea să-şi amintească V.P. Orlov, membru al echipajului sovietic. „Simţeai că stai intr-un butoi de metal pe care cineva îl izbeşte continuu cu un baros.“ Ceea ce iarăşi nu ştiau americanii era că B-59 avea o torpilă nucleară pe care o putea lansa fără să aştepte aprobarea Moscovei. Şi, într-adevăr, căpitanul Saviţki s-a hotărât să lanseze torpila nucleară. Valentin Grigorievici, ofiţerul responsabil de torpilă, a exclamat: „O să murim, dar o să-i scufun dăm pe ei, pe toţi, n-o să facem de râs flota noastră militară!“ Din fericire, decizia de lansare trebuia avizată de trei ofiţeri, iar unul dintre ei, Vasili Arhipov, a spus nu. E trist că foarte puţini au auzit de Arhipov, deşi hotărârea lui a împiedicat pesemne declanşarea celui de al Treilea Război Mondial, şi a fost cea mai preţioasă contribuţie individuală pentru salvarea omenirii din istoria modernă.[38] E cutremurător să te gândeşti ce s-ar fi întâmplat dacă B-59 ar fi fost un submarin autonom controlat de IA, fără oameni în circuitul deciziei.

Două decenii mai târziu, pe 9 septembrie 1983, tensiunea dintre superputeri crescuse din nou: Uniunea Sovietică fusese recent numită „imperiul răului“ de preşedintele american Rónáid Reagan, iar în săptămâna precedentă doborâse un avion de pasageri al companiei Koreán Airlines care se rătăcise în spaţiul ei aerian,

ucigând 269 de oameni, între care un congresman american. Un sistem sovietic automat de avertizare timpurie a raportat că Statele Unite lansaseră asupra Uniunii Sovietice cinci rachete nucleare, lăsând la dispoziţia ofiţerului Stanislav Petrov doar câteva minute pentru a hotărî dacă era o alarmă falsă. Satelitul funcţiona corect, deci, urmând protocolul, trebuia să anunţe un atac nuclear.
Urmându-şi însă instinctul şi socotind că era improbabil ca americanii să atace cu numai cinci rachete, a raportat comandanţilor săi că era o alarmă falsă, fără să ştie că acesta era adevărul. S-a aflat mai târziu că satelitul luase reflexiile Soarelui pe suprafaţa superioară a norilor drept flăcări ale unor motoare de rachetă.[39] Mă întreb ce s-ar fi întâmplat dacă Petrov ar fi fost înlocuit de un sistem LA care să urmeze întocmai protocolul.

Următoarea cursă a înarmărilor?

După cum bănuiţi deja, mă îngrijorează foarte mult sistemele de arme autonome. Dar n-am ajuns încă la principala mea temere: punctul final al unei curse a înarmărilor cu armament LA. în iulie 2or5 mi-am exprimat îngrijorarea în următoarea scrisoare deschisă, alături de Stuart Russell, având susţinerea colegilor mei de la Institutul pentru Viitorul Vieţii:[40]

ARMELE AUTONOME

Scrisoare deschisă din partea cercetătorilor în domeniul inteligenţei artificiale şi roboticii

Armele autonome selectează şi atacă ţinte fără intervenţie umană. Ele ar putea include, de exemplu, quadcoptere†††† înarmate care pot depista şi elimina oameni îndeplinind anumite criterii prestabilite, dar nu

†††† Elicoptere ridicate şi propulsate de patru rotoare. *(N.* t)

includ rachete de croazieră sau drone pilotate de la distanţă, la care oamenii iau toate deciziile de lovire a ţintei Tehnologia IA a ajuns într-un punct în care punerea în funcţiune a unor asemenea sisteme e practic, chiar dacă nu şi legal, posibilă în următorii ani, nu peste decenii, iar miza este enormă: armele autonome reprezintă a treia revoluţie în domeniul militar, după praful de puşcă şi armele nucleare.

S-au adus multe argumente pro şi contra armelor autonome – de pildă, că înlocuirea soldaţilor umani cu maşini e bună prin reducerea numărului victimelor de partea celui care le deţine, dar rea prin scăderea pragului în declanşarea bătăliilor. între- barea-cheie pentru omenire este acum dacă să înceapă o cursă globală a înarmărilor IA sau s-o împiedice. Dacă o mare putere militară porneşte dezvoltarea armei LA, atunci o cursă a înarmărilor e practic inevitabilă, iar punctul final al acestei traiectorii tehnologice este evident armele autonome vor deveni kalaşnikovurile de mâine. Spre deosebire de armele nucleare, ele nu necesită materii prime scumpe sau greu de obţinut, ca atare vor deveni omniprezente, iar pentru marile puteri militare producţia lor în masă va fi ieftină. Va fi doar o chestiune de timp până să apară pe piaţa neagră şi în mâinile teroriştilor, ale dictatorilor care vor să controleze mai bine masele, ale războinicilor dornici de purificare etnică etc. Armele autonome sunt ideale pentru sarcini precum asasinatele, destabilizarea naţiunilor, supunerea populaţiilor şi uciderea selectivă a unui anumit grup etnic. Credem aşadar că o cursă a înarmărilor IA nu va fi în folosul omenirii. Există multe moduri prin care IA poate face câmpurile de bătălie mai sigure pentru oameni, în special pentru civili, fără a crea noi instrumente de ucidere a oamenilor.

Aşa cum majoritatea chimiştilor şi biologilor n-au interesul să construiască arme chimice sau biologice,

majoritatea cercetătorilor LA n-au interesul să construiască arme IA şi nu vor ca alţii să întineze domeniul lor făcând acest lucru, provocând pesemne o puternică reacţie din partea publicului împotriva LA, care să diminueze viitoarele ei beneficii sociale. într-adevăr, chimiştii şi biologii au sprijinit masiv înţelegerile internaţionale care au reuşit să interzică armele chimice şi biologice, la fel cum majoritatea fizicienilor au sprijinit tratatele interzicând plasarea în spaţiu a armelor nucleare şi a armelor laser orbitoare.

Pentru a face mai greu de respins îngrijorările noastre, pe motiv că sunt ale unor ecologişti pacifişti, am vrut ca scrisoarea să fie sem nată de cât mai mulţi cercetători IA şi roboticieni de vârf. Campania Internaţională pentru Controlul Armelor Robotice adunase sute de semnături care cereau interzicerea roboţilor ucigaşi, iar eu speram ca noi sa ne descurcăm şi mai bine. Ştiam că organizaţiile profesionale n-ar fi dispuse să ne ofere listele cu adresele de e-mail ale membrilor lor pentru un scop care putea fi considerat politic, aşa că am încropit liste cu numele cercetătorilor şi instituţii din documente online şi am pus sarcina găsirii adreselor de e-mail pe seama platformei Amazon MTurk. Majoritatea cercetătorilor au adresele de e-mail trecute în site-urile universităţilor lor, iar după 24 de ore şi după ce am plătit 54 de dolari, eram mândrul posesor al unei liste cu adresele a sute de cercetători IA suficient de reputaţi pentru a fi aleşi membri ai Asociaţiei pentru Progresul Inteligenţei Artificiale (Association for the Advancement of Artificial Intelligence – AAAI). Unul dintre ei era profesorul australian Toby Walsh, care a trimis mesaje tuturor celor de pe listă şi a difuzat campania noastră. Participanţii la MTurk din lumea întreagă au produs noi liste pentru Toby, iar curând peste 3 000 de cercetători în IA şi robotică semnaseră scrisoarea noastră deschisă, între care şase foşti preşedinţi AAAI şi lideri ai industriei IA de la Google,

Facebook, Microsoft şi Tesla. O armată de voluntari FLI a lucrat neobosit pentru a valida listele de semnături, eliminând nume trecute în bătaie de joc ca Bill Clinton şi Sarah Connor. S-au mai adăugat alte 17 000 de semnături, între care cea a lui Stephen Hawking, iar după ce Toby a organizat o conferinţă de presă pe această temă la Conferinţa Internaţională Comună a Inteligenţei Artificiale, scrisoarea a intrat în atenţia marilor agenţii de presă din lumea întreagă.

Pentru că biologii şi chimiştii au luat la un moment dat atitudine, domeniile lor sunt acum cunoscute în primul rând pentru crearea medicamentelor şi materialelor folositoare, nu a armelor biologice şi chimice. Acum s-au exprimat şi comunităţile din IA şi robotică: semnatarii scrisorii voiau de asemenea ca domeniile lor să fie cunoscute pentru făurirea unui viitor mai bun, nu a unor noi moduri de a ucide oameni. Dar principala utilizare a IA va fi civilă sau militară? Deşi în acest capitol am vorbit mai mult despre prima, s-ar putea ca în curând să cheltuim mai mulţi bani pentru cea de-a doua – mai ales dacă se declanşează o cursă a înarmărilor IA. Investiţiile în IA din domeniul civil au depăşit în 2016 un miliard de dolari, dar Pentagonul a cerut pentru anul 2017 un buget de 12-15 miliarde de dolari destinaţi proiectelor legate de IA, iar China şi Rusia n-au rămas pesemne indiferente la această declaraţie a lui Robert Work, ministrul adjunct al Apărării: „Vreau ca rivalii noştri să se întrebe se s-o fi ascunzând în spatele cortinei negre.“[41]

Ar trebui semnat un tratat internaţional?

Deşi există acum presiuni internaţionale pentru negocieri privind interzicerea unei anumite forme de roboţi ucigaşi, nu-i limpede ce se va întâmpla, şi s-a declanşat o dezbatere în legătură cu ce anume ar *trebui* să se întâmple. Multe figuri importante din domeniu sunt de

acord că e nevoie ca puterile mondiale să elaboreze un fel de reglementări internaţionale pentru a îndruma cercetările şi utilizarea AWS, dar nu cad de acord cu privire la ce anume trebuie interzis şi cum poate fi impusă o interdicţie. De pildă, ar trebui interzise numai armele autonome letale, sau şi cele care produc răni grave, prin sângerare? Să interzicem proiectarea, producerea sau deţinerea? Interdicţia ar trebui aplicată tuturor sistemelor de arme autonome, sau, aşa cum se spune în scrisoarea noastră, doar celor ofensive, acceptându-se sisteme defensive cum sunt armele autonome antiaeriene şi antirachetă? în acest din urmă caz, AWS trebuie socotite defensive chiar dacă sunt uşor de deplasat în teritoriul inamic? Şi cum poţi aplica un tratat, când majoritatea componentelor armelor autonome au şi o utilizare civilă? De exemplu, nu-i mare diferenţă între o dronă încărcată cu pachete de la Amazon şi una încărcată cu bombe.

Unii participanţi la dezbateri au afirmat că elaborarea unui tratat AWS eficient e extrem de dificilă şi nici n-are rost să încerci. Pe de altă parte, anunţând programul spaţial care avea ca obiectiv Luna, John F. Kennedy a spus că merită să încerci să faci lucruri dificile atunci când succesul va fi important pentru viitorul omenirii. Apoi, mulţi experţi au susţinut că interzicerea armelor biologice şi chimice s-a dovedit benefică, chiar dacă impunerea ei a fost problematică, iar unii au trişat, fiindcă interzicerea a provocat o gravă stigmatizare care a limitat folosirea acelor arme.

În 2016 m-am întâlnit la un dineu cu Henry Kissinger, şi am avut ocazia să-I întreb despre rolul pe care l-a jucat în interzicerea armelor biologice. Mi-a explicat că, pe când era consilier pentru securitatea naţională, l-a convins pe preşedintele Nixon că interdicţia ar fi în folosul securităţii americane. M au impresionat mintea ascuţită şi memoria unui om de 92 de ani, şi m-a fascinat perspectiva

lui de politician. Statele Unite având deja statutul de supraputere graţie forţelor militare convenţionale şi nucleare, aveau mai curând de pierdut decât de câştigat dintr-o cursă mondială a înarmărilor biologice cu rezultat incert. Altfel spus, dacă te afli deja în frunte, e util să urmezi principiul „Dacă nu-i stricat, nu-l repara“. Stuart Russell s-a alăturat conversaţiei noastre, şi am discutat despre aplicarea întocmai a aceluiaşi argument în cazul armelor autonome letale: nu supraputerile ar avea de câştigat dintr-o cursă a înarmărilor, ci micile state lipsite de scrupule şi actorii nonstatali precum teroriştii, care ar avea acces la arme prin piaţa neagră, odată ce ele vor fi fost construite.

Trecându-se la producţia în masă, micile drone ucigaşe echipate cu IA e probabil să coste cu puţin mai mult decât un telefon inteligent. Indiferent dacă un terorist vrea să asasineze un politician ori un bărbat părăsit caută să se răzbune pe fosta lui iubită, e de-ajuns ca ei să încarce fotografia ţintei şi adresa ei în drona ucigaşă: aceasta va zbura spre destinaţie, va identifica şi elimina persoana, iar apoi se va autodistruge, pentru ca nimeni să nu ştie cine e răspunzător. Sau, pentru cei care urmăresc purificarea etnică, aceasta poate fi lesne programată să ucidă doar oameni cu o anumită culoare a pielii ori de o anumită etnie. Stuart prevede că, cu cât vor deveni mai deştepte asemenea arme, cu atât va fi nevoie de mai puţine materiale, muniţii şi bani per crimă. De pildă, se teme de drone de dimensiunea unei albine care vor ucide folosind muniţie minimă împuşcând în ochi, care e suficient de moale pentru a permite unui mic proiectil să pătrundă în creier. Sau s-ar putea agăţa de cap cu nişte gheare metalice, pentru a străpunge apoi craniul cu o încărcătură minusculă. Dacă un milion de asemenea drone ucigaşe pot fi lansate din remorca unui singur camion, avem o înspăimântătoare armă de distrugere în

masă de un tip cu totul nou, care ucide selectiv doar o anumită categorie de oameni.

Un contraargument este că putem elimina aceste griji făcând ca roboţii ucigaşi să fie „etici“ – de pildă, să ucidă numai soldaţii inamici. Dar dacă ni se pare greu să aplicăm o interdicţie, ar fi oare mai uşor să impunem condiţia ca armele autonome să fie roo% etice decât să impunem ca ele să nu fie produse? Şi putem afirma că roboţii ar asculta cu mai multă stricteţe regulile războiului decât soldaţii bine antrenaţi ai naţiunilor civilizate, iar pe de altă parte că dictatorii şi grupările teroriste respectă atât de strict regulile războiului, încât n ar alege niciodată să folosească roboţii intr-un mod care violează aceste reguli?

Războiul cibernetic

Alt aspect militar interesant al IA este că aceasta ţi-ar putea permite să-ţi ataci duşmanul chiar şi fără să construieşti vreo armă, prin război cibernetic. Ca un mic preludiu a ceea ce viitorul ne-ar putea aduce, viermele Stuxnet, despre care se bănuieşte că a fost creat de americani şi israelieni, a infectat centrifugele din programul nuclear iranian de îmbogăţire izotopică şi le-a dezmembrat. Cu cât o societate e mai automatizată şi cu cât devine mai puternic atacul IA, cu atât mai devastator poate fi războiul cibernetic. Dacă poţi virusa şi distruge automobilele care se conduc singure ale inamicului, avioanele lui autopilotate, reactoarele nucleare, roboţii industriali, sistemele financiar şi de comunicaţii, reţelele electrice, atunci îi poţi efectiv distruge economia şi paraliza apărarea. Dacă îi poţi virusa şi sistemele militare, cu atât mai bine.

Am început acest capitol arătând cât de benefică poate fi pentru omenire, în viitorul apropiat, IA – cu condiţia să devină robustă şi imună Ia atacurile informatice. Deşi IA

însăşi poate fi folosită pentru a face sistemele IA mai robuste, sprijinind astfel apărarea în războiul cibernetic, ea poate sprijini şi agresiunea. Garantarea faptului că apărarea predomină trebuie să fie unul dintre obiectivele pe termen scurt ale dezvoltării IA – altminteri întreaga tehnologie pe care o construim se poate întoarce împotriva noastră!

JOBURI ŞI SALARII

Până acum acest capitol s-a referit în primul rând la impactul IA asupra noastră în calitate de *consumatori,* oferindu-ne noi produse şi servicii la preţuri accesibile. Dar cum ne va influenţa în calitate de *muncitori,* transformând piaţa muncii? Dacă găsim o cale de a ne spori prosperitatea prin automatizare, fără a-i priva pe oameni de venituri sau scopuri, atunci ne putem făuri un viitor fantastic, cu un timp liber şi o bogăţie nemaiîntâlnite pentru toată lumea. Puţini s-au gândit la acest aspect mai îndelungat şi mai profund decât economistul Erik Brynjolfsson, colegul meu de la MIT. Deşi e mereu îngrijit şi impecabil îmbrăcat, ascendenţa sa islandeză mă face să-mi închipui că abia de curând şi-a ras barba roşie de viking şi şi-a tuns chica pentru a se alătura şcolii noastre de afaceri. Fără îndoială însă nu s-a debarasat şi de ideile lui îndrăzneţe, iar viziunea sa optimistă asupra pieţei muncii a numit-o „Atena Digitală". Motivul principal pentru care cetăţenii atenieni din Antichitate au dus vieţi confortabile, în care se puteau bucura de democraţie, artă şi jocuri, a fost că aveau sclavi care să facă mare parte din muncă. De ce să nu înlocuim deci sclavii cu roboţi echipaţi cu IA, creând o utopie digitală de care toată lumea să se bucure? Nu numai că ar elimina stresul şi corvoada şi ar produce tot ce ne dorim azi, dar economia condusă de IA, în viziunea lui Erik, ne-ar oferi şi o mulţime de noi produse şi servicii pe care

consumatorii de azi nici măcar nu şi-au dat seama că şi le doresc.

Tehnologie şi inegalitate

Putem ajunge la Atena Digitală a Iui Erik dacă salariul pe oră al fiecăruia continuă să crească an de an, aşa încât cei care vor mai mult timp liber să poată treptat munci mai puţin, îmbunătăţindu-şi constant nivelul de trai. Figura 3.5 arată că tocmai asta s-a întâmplat în Statele Unite de la Al Doilea Război Mondial până pe la mijlocul anilor ’70: deşi a existat o inegalitate a veniturilor, dimensiunea totală a tortului a crescut aşa încât aproape toţi au primit o felie mai mare. Apoi însă, după cum recunoaşte şi Erik, ceva s-a schimbat: figura 3.5 arată că deşi economia a continuat să crească, sporind venitul mediu, câştigurile din ultimele patru decenii s-au îndreptat către cei bogaţi, în special către cei mai bogaţi 1 %, iar la cei mai săraci 90% veniturile au stagnat. Accentuarea inegalităţii reiese şi mai clar dacă ne uităm nu la venit, ci la nivelul bunăstării. Pentru cele mai sărace 90% dintre familiile americane, activul net mediu a fost de circa 85 000 de dolari în 2012 – acelaşi ca în urmă cu douăzeci de ani –, în vreme ce pentru cele mai bogate 1% s-a dublat în aceeaşi perioadă, ajungând la r 4 milioane de dolari (luând în calcul inflaţia).[41] Pe plan mondial diferenţele sunt şi mai mari, în 2013, averea însumată a jumătăţii mai sărace din populaţia lumii (peste 3,6 miliarde de oameni) este egală cu averea celor mai bogaţi opt oameni.[43] La conferinţa noastră din Puerto Rico, Erik a afirmat că progresul inteligenţei artificiale şi automatizării va face să crească tortul economiei, dar nici o lege economică nu spune că toată lumea, sau măcar majoritatea, va avea de câştigat.

Deşi economiştii cad de acord asupra faptului că inegalitatea se accentuează, există o interesantă controversă privind cauzele fenomenului şi tendinţele viitoare. Cei cu orientări politice de stânga susţin că

principalele cauze sunt globalizarea şi/sau politicile economice precum reducerile fiscale pentru cei bogaţi. Erik Bryn- jolfsson şi colaboratorul său de la MIT Andrew McAfee susţin însă că alta e principala cauză: tehnologia.[44] Mai precis, tehnologia digitală acţionează în trei moduri asupra inegalităţii.

în primul rând, prin înlocuirea vechilor joburi cu unele care necesită mai multe aptitudini, tehnologia i-a recompensat pe cei educaţi: de la mijlocul anilor ’70, salariile absolvenţilor de facultate au crescut cu circa 25%, iar salariile celor care au abandonat şcoala au scăzut cu circa 3O%.[45]

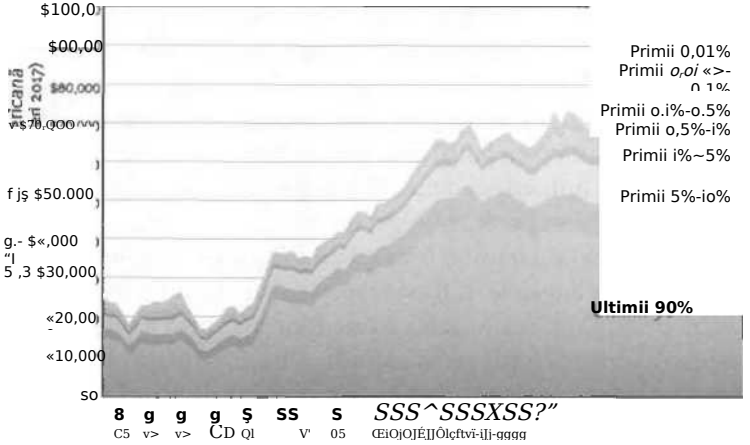


Figura 3.5 Creşterea venitului mediuîn ultimul secol ţi partea din acesta care a revenit diferitelor grupuri. înainte de anii'70, bogaţi şi săraci au prosperat în ace la ţi ritm, dar apoi majoritatea câştigurilor s-au îndreptat către primii 1%, iar ultimii 90% n-au câştigat în medie aproape nimic.[46] Sumele au fost corectate în raport cu inflatia ti reprezintă dolari din 2017.

în al doilea rând, începând din anul 2000, o parte tot mai mare din venitul corporaţiilor s-a îndreptat către cei care deţin companiile, nu către cei care muncesc în ele – iar atâta timp cât automatizarea continuă, este de aşteptat ca aceia care deţin maşinile să capete o felie tot mai mare din tort. Raportul dintre capital şi muncă poate juca un rol foarte important pentru economia digitală în

ascensiune, pe care Nicholas Negroponte, un profet al tehnologiei, o defineşte prin mişcarea biţilor, nu a atomilor. Acum când totul, de la cărţi la filme şi la instrumentele pentru întocmirea declaraţiilor fiscale, a devenit digital, copiile pot fi vândute pretutindeni în lume cu costuri practic nule, fără angajarea unui personal suplimentar. Aceasta permite ca majoritatea profitului să revină investitorilor, nu muncitorilor, şi explică de ce profitul însumat al primelor trei companii din Detroit (General Motors, Ford şi Chrysler) în 1990 a fost aproape identic cu profitul însumat al primelor trei companii din Silicon Valley (Google, Apple, Facebook) în 2014, acestea din urmă având de nouă ori mai puţini angajaţi şi valorând de treizeci de ori mai mult la bursă.[47]

în al treilea rând, Erik şi colaboratorii săi susţin că economia digitală favorizează adesea superstarurile în detrimentul tuturor celorlalţi. J.K. Rawling, autoarea Iui Harry Potter, a devenit primul scriitor miliardar, şi s-a îmbogăţit mult mai mult decât Shakespeare fiindcă poveştile ei au putut fi transmise sub formă de text, filme şi jocuri către miliarde de oameni la un preţ foarte scăzut. în mod asemănător, Scott Cook a câştigat un miliard cu software-ul pentru întocmirea declaraţiilor fiscale TurboTax, care, spre deosebire de persoanele angajate în acest scop, poate fi vândut şi descărcat pe intemet. Cum majoritatea oamenilor n-ar prea cumpăra al zecelea asemenea software în ordinea calităţii, pe piaţă e loc doar pentru puţine superstaruri. Asta înseamnă că dacă toţi părinţii din lume îşi sfătuiesc copiii să devină următoarea sau următorul J.K. Rawling, Gisele Bundchen, Matt Damon, Cristiano Ronaldo sau Elon Musk, aproape nici unul dintre copii nu va găsi că e o strategie viabilă pentru cariera de urmat.

Sfaturi profesionale pentru copii

Ce carieră *ar trebui* să-i sfătuim pe copiii noştri să

urmeze? Eu i-am încurajat pe ai mei să caute profesiuni în care maşinile deocamdată nu se descurcă, fiind deci improbabil să se automatizeze în viitorul apropiat. Previziuni recente pentru momentul când diferite joburi vor fi preluate de maşini scot în evidenţă câteva întrebări pe care e util să ţi le pui înainte de a începe educaţia pentru a urma o anumită carieră.[48] De exemplu:

- Presupune interacţiunea cu oamenii şi folosirea inteligenţei sociale?
- Implică oare creativitatea şi găsirea unor soluţii ingenioase?
- Presupune munca într-un domeniu impredictibil?

Cu cât poţi răspunde prin da la mai multe dintre aceste întrebări, cu atât e probabil să fie mai bună alegerea carierei Asta înseamnă că profesiuni relativ sigure sunt cele de profesor, asistentă medicală, doctor, dentist, om de ştiinţă, antreprenor, inginer, avocat, asistent social, cleric, artist, coafeză sau kinetoterapeut.

Dimpotrivă, joburi care presupun acţiuni repetitive sau structurate, în condiţii predictibile, este probabil să devină nu peste mult timp automatizate. Computerele şi roboţii industriali au preluat de mult cele mai simple asemenea joburi, iar tehnologia în curs de perfecţionare elimină multe altele, de la agent de vânzări la lucrător în depozit, casier, feroviar, brutar şi bucătar necalificat.[49] Şoferii de camioane, autobuze, taxiuri şi maşini Uber/ Lyft e probabil să urmeze în curând. Există numeroase alte profesii (între care asistent juridic, analist de credite şi contabil) care, deşi nu sunt încă ameninţate cu dispariţia completă, îşi văd deja automatizate majoritatea sarcinilor, prin urmare necesită mai puţini oameni.

Dar evitarea automatizării nu e singura provocare legată de carieră. în această epocă digitală globală, a dori să devii scriitor profesionist, regizor de film, actor, atlet

sau creator de modă e riscant din alt motiv: deşi oamenii din aceste profesii nu vor fi concuraţi serios de maşini în viitorul apropiat, ei vor fi tot mai puternic concuraţi de alţi oameni de pretutindeni, conform teoriei superstarului menţionată mai sus, iar foarte puţini vor reuşi.

în multe cazuri, e imprecis să dai sfaturi Ia nivelul unui întreg domeniu: numeroase joburi, iară a fi complet eliminate, vor avea multe dintre sarcinile lor automatizate. De pildă, dacă vrei să studiezi medicina, nu deveni acel radiolog care analizează imaginile şi e înlocuit de calculatorul IBM Watson, ci doctorul care prescrie analiza radiologică, discută rezultatele cu pacientul şi hotărăşte tratamentul. Dacă vrei să intri în finanţe, nu deveni acel analist cantitativ care aplică algoritmi asupra datelor şi e înlocuit de un software, ci administrator de fonduri care foloseşte rezultatele analizei cantitative pentru a lua decizii strategice de investiţii. Dacă vrei să intri în domeniul juridic, nu deveni acel asistent care parcurge mii de acte pentru faza de documentare şi e eliminat prin automatizare, ci avocatul care îşi sfătuieşte clientul şi pledează la tribunal.

Am examinat deocamdată ce pot face indivizii pentru a-şi maximiza succesul pe piaţa muncii în epoca IA. Dar ce pot face guvernele pentru a veni în sprijinul forţei de muncă? De pildă, ce sistem de învăţământ oferă cea mai bună pregătire pentru o piaţă a muncii în care IA evoluează rapid. Este în continuare actualul nostru model cu două decenii de învăţământ, urmate de patru decenii de muncă specializată? Sau e de preferat ca oamenii să lucreze câţiva ani, să se întoarcă la studii pentru un an, apoi să lucreze iar câţiva ani?[50] Sau continuarea educaţiei (eventual online) să fie regula oricărui job?

Ce politici economice sunt cele mai utile pentru crearea unor joburi noi profitabile? Andrew McAfee susţine că există destule asemenea politici, între care

investiţii masive în cercetare, educaţie şi infrastructură, facilitarea migraţiei şi stimularea antrepre- noriatului. El crede că „playbook-ul Econ roi oferă sfaturi clare, dar nu e urmat“, cel puţin în Stele Unite.[51]

Vor deveni oamenii în cele din urmă neangajabili?

Ce se va întâmpla dacă IA continuă să se perfecţioneze, automatizând tot mai multe joburi? Mulţi oameni rămân optimişti, afirmând că joburile automatizate vor fi înlocuite de unele noi, mai bune. în fond, aşa s-a întâmplat mereu până acum, de când, în timpul Revoluţiei Industriale, pe ludiţi i-a îngrijorat şomajul tehnologic.

Alţii sunt însă pesimişti şi susţin că de data asta e altfel, iar tot mai mulţi oameni vor deveni nu doar şomeri, ci neangajabili.[52] Pesimiştii afirmă că piaţa liberă stabileşte salariile pe baza cererii şi ofertei, şi că oferta sporită de muncă automată ieftină va duce în cele din urmă la scăderea salariilor oamenilor mult sub costul vie- ţii. Din moment ce salariul pe piaţă pentru un job e dat de cât costă pe oră oricine sau orice efectuează cel mai ieftin acea muncă, salariile au scăzut de fiecare dată când a fost posibil să extemalizezi o anumită ocupaţie către o ţară mai săracă sau o maşină ieftină. în timpul Revoluţiei Industriale, am început să ne gândim cum să înlocuim muşchii noştri cu maşini, iar oamenii s-au orientat spre joburi mai bine plătite, în care îşi foloseau mai mult mintea. „Gulerele albastre“ au fost înlocuite de „gulerele albe“. Acum ne gândim cum să înlocuim minţile noastre cu maşini. Dacă vom reuşi, atunci ce joburi rămân pentru noi?

Unii optimişti susţin că, după joburile fizice şi mentale, următorul *boom* va fi în joburile *creative,* dar pentru pesimişti creativitatea nu-i decât un proces mental, aşa încât şi ea va fi deprinsă în cele din urmă de IA. Alţi optimişti speră ca următorul boom să fie în noi profesii apărute graţie tehnologiei, pe care nici nu ni le-am închipuit până acum. Cine şi-ar fi imaginat în timpul Revoluţiei Industriale că urmaşii lui vor lucra într-o bună zi ca web designezi sau ca şoferi Uber? Pentru pesimişti însă, acestea sunt iluzii deşarte, nesusţinute de datele empirice. Ei subliniază că acelaşi raţionament l-am fi putut face cu un secol în urmă, înaintea revoluţiei informatice, şi am fi prevăzut că majoritatea profesiilor de azi vor fi fiind unele noi şi nebănuite, apărute graţie tehnologiei. Această predicţie ar fi fost eronată, după cum se vede

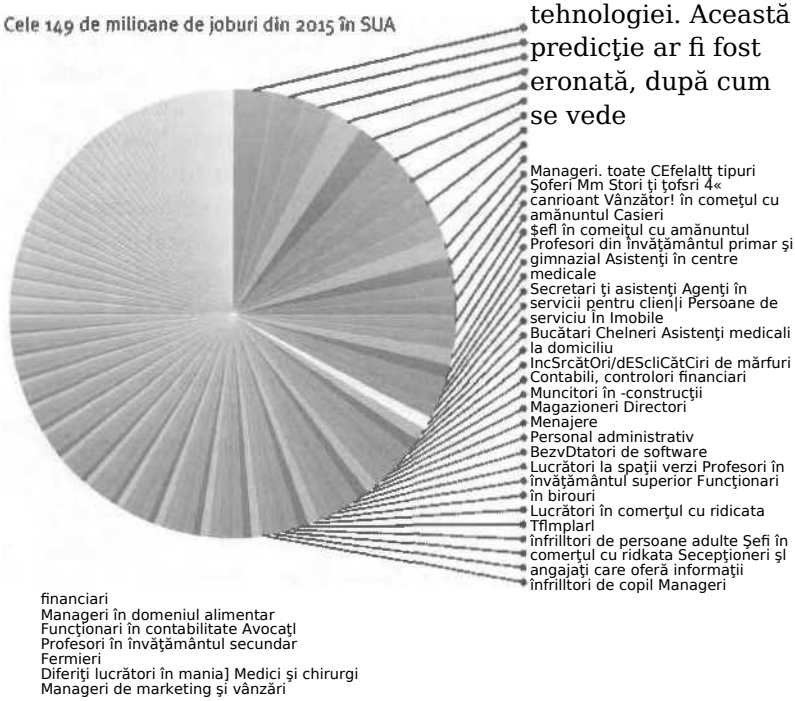


Figura 3.6 Graficul reprezintă ocupaţiile celor 149 de milioane de americani care aveau un jobîn 2015, cu cele 535 de categorii de joburi conform Departamentului de Statistică a Muncii, clasificate în

ordinea descrescătoare a numărului de lucrători.53 Sunt marcate toate ocupaţiile cu peste un milion de lucrători. Până la poziţia 21 nu există noi ocupaţii create de tehnologia informaticii. Figura e inspirată de o analiză a lui Federico Pistono.[54] în figura 3.6: marea majoritate a ocupaţiilor de azi sunt cele care existau deja cu un secol în urmă, iar dacă le clasificăm după numărul de joburi pe care le oferă, ajungem la poziţia 21 de pe listă pentru a găsi o nouă ocupaţie: dezvoltatori de software, care reprezintă mai puţin de 1% din piaţa muncii în Statele Unite.

înţelegem mai bine ce se întâmplă dacă ne amintim de figura 2.2 din capitolul 2, care prezintă peisajul inteligenţei umane, altitudinea exprimând cât de dificil le este maşinilor să execute diferite sarcini, iar nivelul în creştere al apei indicând ce pot face acum maşinile. Tendinţa principală a pieţei muncii nu e de orientare către profesii cu totul noi. Ne înghesuim în acele zone de uscat din figura 2.2 care n-au fost încă inundate de mareea tehnologiei! Figura 3.6 arată că ele nu formează o singură insulă, ci un arhipelag complex, cu insuliţe şi atoluri corespunzând tuturor acelor lucruri pe care maşinile nu le pot face deocamdată cu un cost la fel de mic ca oamenii. E vorba nu doar de profesii de înaltă tehnologie, ca dezvoltator de software, dar şi de o gamă largă de joburi implicând un nivel tehnologic scăzut, care se bazează pe dexteritatea şi aptitudinile noastre sociale superioare, de la cel de kinetoterapeut la cel de actor. E oare posibil ca IA să ne eclipseze atât de rapid în îndeplinirea sarcinilor intelectuale, încât ultimele joburi care ne rămân să fie cele de nivel tehnologic scăzut? Un prieten mi-a zis în glumă că probabil ultima meserie care ne rămâne va fi chiar prima meserie apărută în istorie: prostituţia. Apoi însă i-a spus asta unui robotician japonez, care a protestat: „Nu, roboţii se pricep foarte bine la asemenea lucruri!“

Pesimiştii afirmă că deznodământul e evident: întregul arhipelag va fi scufundat, şi nu vor mai exista joburi in care oamenii să muncească mai ieftin decât maşinile. în

cartea sa din 2007 *Farewell to Alms,* economistul Gregory Clark arată că putem afla câte ceva despre viitorul nostru dintr-o comparaţie cu cabalinele, închipuiţi-vă doi cai asistând pe la 1900 la începuturile automobilului şi meditând asupra viitorului lor.

> „Mă îngrijorează şomajul tehnologic.“
> „Hai, nu fi ludit. Strămoşii noştri au spus acelaşi lucru atunci când motoarele cu aburi ne-au luat joburile din industrie şi trenurile ne-au luat jobul de a trage poştalioanele. Avem azi mai multe joburi ca niciodată, şi sunt şi mai bune: prefer să trag o trăsură uşoară prin oraş decât să merg ziua întreagă in cerc ca să acţionez o pompă stupidă dintr-o mină.“ „Dar dacă motoarele cu ardere internă se dezvoltă?“
> „Sunt sigur că vor fi joburi noi pentru cai, pe care încă nu ni le imaginăm. Aşa s-a întâmplat de fiecare dată până acum, la fel ca atunci când s-au inventat roata şi plugul.“

Din păcate, acele joburi încă neimaginate pentru cai n-au mai apărut niciodată. Caii de care nu era nevoie au fost sacrificaţi, aşa încât şeptelul cabalin din Statele Unite s-a prăbuşit de la circa 26 de milioane în 1915 la circa 3 milioane în 1960.[55] Dacă muşchii mecanici au făcut să devină inutili caii, minţile mecanice vor face oare acelaşi lucru cu oamenii?

Venituri fără joburi

Cine are aşadar dreptate, cei care susţin că joburile automatizate vor fi înlocuite de unele mai bune, sau cei care susţin că majoritatea oamenilor vor deveni neangajabili? Dacă progresul IA continuă neabătut, atunci *ambele* tabere ar putea avea dreptate: una pe termen scurt, cealaltă pe termen lung. Insă deşi oamenii vorbesc adesea despre dispariţia joburilor în termeni apocaliptici, s-ar putea să nu fie neapărat un lucru rău! Ludiţii erau

obsedaţi de anumite meserii, ignorând posibilitatea ca altele să aibă aceeaşi valoare socială. La fel, s-ar putea ca cei obsedaţi azi de anumite meserii să fie prea mărginiţi: vrem joburi fiindcă ele ne oferă un venit şi un scop, însă dată fiind bogăţia de resurse produse de ma şini, ar trebui să putem găsi alte căi de a ne asigura venituri şi scopuri *fără* joburi. Ceva asemănător s-a întâmplat în povestea cabalină, care nu s-a încheiat cu dispariţia tuturor cailor. De fapt, numărul cailor s-a triplat din 1960 încoace, căci ei au fost protejaţi de un soi de sistem de asistenţă socială cabalină: deşi aceştia nu-şi puteau plăti facturile, oamenii s-au hotărât să aibă grijă de cai, păstrându-i pentru divertisment, sport şi companie. Putem oare avea şi noi grijă de semenii noştri umani aflaţi în suferinţă?

Să începem cu problema venitului: redistribuirea fie şi a unei mici părţi din tortul economic în creştere ar permite oricui s-o ducă mai bine. Mulţi susţin nu doar că o *putem* face, dar *trebuie s-o* facem. La întrunirea din 2016, unde Moshe Vardi a vorbit despre imperativul moral de a salva vieţi cu tehnologia IA, eu am afirmat că tot un imperativ moral este şi să pledăm pentru folosirea ei benefică, incluzând aici distribuirea bogăţiei. Erik Brynjolfsson, participant şi el la întrunire, a spus: „Dacă cu toată această nouă generare a bogăţiei nu putem împiedica nici măcar ca jumătate din omenire s-o ducă mai rău, atunci să ne fie ruşine!“

Există diferite propuneri pentru distribuirea bogăţiei, fiecare cu partizanii şi criticii ei. Cea mai simplă este aceea privind *venitul de bază,* conform căreia toţi oamenii primesc o plată lunară, fără precondiţii sau cerinţe. Câteva experimente la scară mică sunt plănuite sau se află în curs de desfăşurare în Canada, Finlanda şi Olanda. Promotorii lor susţin că un venit de bază e mai eficient decât alte soluţii cum ar fi plata ajutoarelor sociale pentru cei nevoiaşi, fiindcă elimină disputele administrative

legate de cine merită şi cine nu. Plata ajutoarelor sociale conform necesităţilor a fost criticată şi pentru că descurajează munca, ceea ce devine însă irelevant într-un viitor fără joburi, în care nimeni nu munceşte.

Guvernele îşi pot ajuta cetăţenii nu doar dându-le bani, ci şi oferindu-le gratuit sau sponsorizat servicii precum drumuri, poduri, parcuri, transport public, îngrijirea copiilor, educaţie, asistenţă medicală, cămine pentru pensionari şi acces la intemet; de fapt, multe guverne oferă deja majoritatea acestor servicii. Spre deosebire de venitul de bază, asemenea servicii finanţate de guvern îndeplinesc două obiective: reduc costul vieţii şi creează joburi. Chiar şi într-un viitor în care maşinile îi depăşesc pe oameni în toate meseriile, guvernele pot alege să-i plătească pe oameni pentru a îngriji copii şi bătrâni, în loc să-i lase pe roboţi s-o facă.

Progresul tehnologic poate face ca numeroase produse sau servicii importante să devină gratuite chiar şi fără intervenţia guvernului. De exemplu, oamenii plăteau înainte pentru enciclopedii, atlase, expedierea scrisorilor sau convorbiri telefonice, dar acum oricine e conectat la intemet are acces gratuit la toate acestea, precum şi la videoconferinţe, partajarea fotografiilor, medii de socializare, cursuri online şi nenumărate alte noi servicii. Numeroase alte lucruri care pot fi importante pentru o persoană cum ar fi un tratament cu antibiotice care îţi salvează viaţa – au devenit extrem de ieftine. Aşadar, graţie tehnologiei, chiar şi mulţi oameni săraci au astăzi acces la lucruri care Ie lipseau în trecut celor mai bogaţi. Unii deduc de aici că venitul necesar unei vieţi decente scade.

Dacă maşinile pot oferi într-o bună zi toate bunurile şi serviciile actuale cu un cost minim, atunci există suficientă bogăţie pentru ca toţi s-o ducă mai bine. Cu alte cuvinte, până şi impozite relativ modeste ar putea permite

guvernelor să plătească un venit de bază şi servicii gratuite. Dar faptul că distribuirea bogăţiei *poate* avea loc, nu înseamnă şi că *va* avea loc, iar acum în mediul politic se desfăşoară dezbateri aprige în jurul întrebării dacă *trebuie* să aibă loc. După cum am văzut, tendinţa actuală în Statele Unite pare să meargă în sens contrar, unele grupuri devenind tot mai sărace decenii la rând. Deciziile politice privind distribuirea bogăţiei societăţii vor avea efecte asupra tuturor, prin urmare la discuţia legată de tipul de economie pe care vrem s-o făurim în viitor trebuie să participe toată lumea, nu doar cercetătorii IA, roboti- cienii şi economiştii.

Mulţi susţin că reducerea inegalităţii veniturilor este o idee bună nu doar în viitorul dominat de IA, dar şi azi. Deşi argumentul principal este de regulă de ordin moral, există şi dovezi că reducerea inegalităţii face ca democraţia să funcţioneze mai bine: dacă există o clasă de mijloc educată, electoratul e mai greu de manipulat, şi e mai dificil pentru un număr mic de oameni sau companii să cumpere o influenţă necuvenită asupra guvernului. La rândul ei, o democraţie mai bună e favorabilă unei economii mai bine administrate, mai puţin coruptă, mai eficientă, cu o creştere mai rapidă, spre folosul tuturor.

Scopuri fără joburi

Joburile oferă oamenilor nu doar bani. în 1759, Voltaire spunea că „munca ne fereşte de trei mari rele: plictiseala, viciul şi sărăcia“. Invers, a da oamenilor un venit nu-i de ajuns pentru a garanta starea lor de bine. împăraţii romani ofereau atât pâine, cât şi circ pentru a-şi mulţumi supuşii, iar Isus spunea „Nu numai cu pâine va trăi omul“. Care sunt deci acele lucruri de preţ pe care le aduc joburile, în afară de bani, şi în ce fel poate o societate fără joburi să le ofere?

Răspunsurile la aceste întrebări sunt extrem de complicate, fiindcă unii oameni îşi detestă joburile, iar

alţii Ie îndrăgesc. Apoi, sunt copii, studenţi şi gospodine care se simt minunat fără joburi, în vreme ce istoria e plină de poveşti despre urmaşi şi prinţi răsfă- ţaţi care se cufundă în plictis şi depresie. O analiză din 2012 arăta că şomajul tinde să aibă efecte negative pe termen lung asupra stării de bine, în vreme ce pensionarea are aspecte atât pozitive, cât şi negative.[56] Domeniul în plină dezvoltare al *psihologiei pozitive* a identificat câteva elemente care îţi întăresc sentimentul că trăieşti bine şi ai un scop în viaţă, şi a găsit că unele joburi (dar nu toate!) îţi pot oferi multe dintre aceste elemente, de exemplu:[57]

- o reţea socială de prieteni şi colegi
- un stil de viaţă sănătos şi moral
- respect, stimă de sine, auto-eficienţă şi senzaţia plăcută de „fluiditate", născute din faptul că faci ceva la care te pricepi
- sentimentul că e nevoie de tine şi că eşti deosebit
- sentimentul că ai un rost dat de faptul că aparţii unei entităţi mai vaste decât tine şi o slujeşti

Asta ne dă motive de optimism, căci toate aceste lucruri pot fi obţinute şi în afara locului de muncă – de pildă, prin sport, hobby-uri şi învăţare, împreună cu familia, prietenii, în echipe, cluburi, grupuri comunitare, şcoli, organizaţii religioase şi umaniste, mişcări politice şi alte instituţii. Pentru a făuri o societate cu un grad scăzut de ocupare a forţei de muncă, care să nu degenereze într-un comportament autodistructiv, ci să prospere, este nevoie deci să înţelegem cum putem stimula aceste activităţi care induc starea de bine. Acest demers trebuie să-i implice nu doar pe oamenii de ştiinţă şi pe economişti, dar şi pe psihologi, sociologi şi educatori. Dacă se fac eforturi serioase vizând starea de bine a tuturor, bazate pe o parte din avuţia generată de viitoarea IA, atunci societatea ar putea prospera ca niciodată în trecut. Fiecare ar trebui să

se simtă fericit ca şi cum şi-ar fi găsit jobul visat, iar odată scăpat de constrângerea de a munci pentru bani, nu mai există limite.

INTELIGENŢĂ LA NIVEL UMAN?

Am văzut în acest capitol că IA ne poate îmbunătăţi mult viaţa în viitorul apropiat, cu condiţia să anticipăm corect şi să evităm diversele capcane. Dar cum stau lucrurile pe termen lung? Se va opri oare progresul IA din pricina unor obstacole insurmontabile, sau vor reuşi până Ia urmă cercetătorii IA să-şi îndeplinească obiectivul iniţial de a construi o inteligenţă artificială generală la nivel uman? Am văzut în capitolul precedent că legile fizicii permit unor grămezi de materie cu anumite proprietăţi să memoreze, să calculeze şi să înveţe, şi că nu interzic unor asemenea grămezi s-o facă într-o bună zi cu mai multă inteligenţă decât grămezile de materie din capetele noastre. Dacă sau când vom reuşi noi, oamenii, să construim această IAG supraumană nu-i la fel de limpede. Am văzut în primul capitol că deocamdată nu putem şti asta – opiniile specialiştilor de frunte din IA sunt împărţite, cei mai mulţi dintre ei fac estimări care se întind pe un spectru larg, de la decenii la secole, iar unii cred că acest lucru nu se va întâmpla niciodată. Previziunea e dificilă fiindcă, atunci când explorezi un teritoriu necartografiat, nu ştii câţi munţi te despart de destinaţie. De regulă, îi vezi doar pe cei din apropiere, şi trebuie să-i escaladezi pentru a descoperi următorul obstacol.

Ce se poate întâmpla în viitorul imediat? Chiar dacă am cunoaşte cea mai bună metodă de a construi o IAG la nivel uman folosind hardware-ul calculatoarelor de azi, ceea ce nu e cazul, am avea nevoie de o mare cantitate de hardware pentru a obţine puterea de calcul brută necesară. Care este deci puterea de calcul a unui creier

uman, măsurată în biţi şi FLOFS (vezi capitolul 2)?‡‡‡‡ E o întrebare delicată, iar răspunsul depinde critic de felul în care o formulăm:

- întrebarea 1: De câţi FLOPS e nevoie pentru a simula un creier uman?
- întrebarea 2; De câţi FLOPS e nevoie pentru inteligenţa umană?
- întrebarea 3: Câţi FLOPS poate efectua un creier uman?

S-au publicat o mulţime de articole pe tema întrebării 1, iar răspunsul tipic este de ordinul sutei de petaFLOPS, adică 10^{17} FLOPS.[58] Este cam aceeaşi putere de calcul ca aceea a lui Suway TaihuLight (figura 3.7), în 2016 cel mai rapid supercomputer, care a costat vreo 300 de milioane de dolari. Chiar dacă am şti cum să-l folosim pentru a simula creierul unui muncitor înalt calificat, singurul beneficiu ar fi că am avea simularea muncii acelui om, dacă am putea închiria TaihulLight la un preţ mai mic decât costul lui pe oră. Ba chiar ar trebui să plătim mai mult, fiindcă numeroşi oameni de ştiinţă cred că, pentru a replica precis inteligenţa unui creier, nu-l putem trata în modelul de reţea neuronală simplificat matematic din capitolul 2. Poate că ar trebui să-l simulăm la nivelul moleculelor individuale, sau chiar la nivelul particulelor subatomice, ceea ce ar necesita incomparabil mai mulţi FLOPS.

Răspunsul la întrebarea 3 este mai simplu: sunt foarte slab la înmulţirea numerelor cu 19 cifre, şi mi-ar lua mult timp chiar dacă îmi dai un creion şi o hârtie. Aş ajunge la mai puţin de o,or FLOPS, cu 19 ordine de mărime sub răspunsul la întrebarea 1! Motivul acestei uriaşe

‡‡‡‡ Reamintesc că FLOPS sunt operaţii în virgulă mobilă pe secundă – de pildă, cât de multe numere de 19 cifre pot fi înmulţite pe secundă. (N. a.)

discrepanţe este că supercomputerele şi creierele sunt optimizate pentru sarcini extrem de diferite. O discrepanţă similară apare la aceste întrebări:

Figura 37 TaihuLight, în 2016 cel mai rapid supercomputer, a cărui putere de calcul s-ar putea s-o depăşească pe cea a creierului uman.

Cât de bun e un tractor într-o cursă de Formula 1?
Cât de bine poate face o maşină de Formula 1 munca unui tractor?

Prin urmare, la care din aceste două întrebări privind FLOPS încercăm să răspundem pentru a prezice viitorul IA? La nici una! Dacă am vrea să simulăm creierul uman, ne-ar preocupa întrebarea 1, dar pentru a construi IAG la nivel uman cea care contează este întrebarea 2. Nimeni nu cunoaşte deocamdată răspunsul la ea, însă ar putea fi mult mai uşor decât să simulezi un creier, dacă fie adaptăm software-ul pentru a se potrivi mai bine calculatoarelor de azi, fie construim un hardware mai asemănător creierului (s-au făcut mari progrese cu aşa-numitele cipuri neuromorfe).

Hans Moravec a estimat răspunsul comparând ce este de corn parat pentru un calcul pe care atât creierele, cât şi calculatoarele de azi îl pot executa eficient: anumite sarcini de procesare de nivel scăzut a imaginii, pe care retina umană le efectuează în spatele globului ocular, şi ale cărei rezultate le trimite apoi creierului prin nervul optic.[59] A făcut socoteala că replicarea calculelor unei retine pe un computer convenţional necesită aproximativ un miliard de FLOPS, şi că întregul creier efectuează de aproximativ zece mii de ori mai multe calcule decât o

retină (pe baza comparaţiei între volumele şi numerele de neuroni), aşa încât capacitatea de calcul a creierului este de circa 10^{13} FLOPS – aproximativ puterea unui computer optimizat de o mie de dolari în 2015!

în concluzie, nu avem garanţia absolută că vom reuşi să construim o IAG la nivel uman în timpul vieţii noastre – sau vreodată. Dar nu există nici un argument imbatabil că nu am avea suficientă putere a hardware-ului sau că ar fi prea scump. Nu ştim cât de departe ne aflăm de linia de finiş în privinţa arhitecturilor, algoritmilor şi software-ului, dar progresul actual e rapid, iar la provocări răspunde o comunitate globală în continuă creştere de cercetători IA talentaţi. Cu alte cuvinte, nu putem respinge posibilitatea ca IAG să atingă în cele din urmă nivelul uman şi să-l depăşească. Ca atare, în următorul capitol vom analiza această posibilitate şi vom încerca să vedem unde ne poate conduce ea.

REZUMAT:

• Progresul IA pe termen scurt ne poate îmbunătăţi mult vieţile pe nenumărate căi, de la sporirea eficienţei vieţilor noastre personale, reţelelor electrice şi pieţelor financiare, până la salvarea de vieţi omeneşti prin automobile care se conduc singure, roboţi chirurgicali şi sisteme IA de diagnosticare,

• Când permitem sistemelor din lumea reală să fie controlate de IA, e esenţial să învăţăm cum poate deveni mai robustă IA, făcând ceea ce vrem noi să facă. Aceasta presupune rezolvarea unor probleme tehnice dificile legate de verificare, validare, securitate şi control.

• Această nevoie de robusteţe sporită e presantă mai ales pentru sistemele de arme controlate de IA, pentru care miza poate fi uriaşă.

• Mulţi cercetători IA şi roboticieni de vârf au pledat în favoarea unui tratat internaţional care să interzică anumite tipuri de arme autonome, pentru a evita o cursă a înarmărilor scăpată de sub control, care ar pune la dispoziţia oricui are portofelul plin şi vrea să-şi plătească nişte poliţe maşini de ucis avantajoase.

• IA poate face mai corect şi mai eficient sistemul nostru juridic, cu condiţia ca robojudecătorii să fie transparenţi şi nepărtinitori.

• Legile noastre trebuie urgent actualizate pentru a ţine pasul cu IA, care pune dificile probleme juridice, Implicând confidenţialitatea, responsabilitatea şi reglementarea.

• Până să ne facem griji că maşinile inteligente ne-ar putea înlocui complet, ele ne vor fi luat în mare măsură locul pe piaţa muncii.

■ Acesta nu e neapărat un lucru rău, câtă vreme societatea redistribuia o parte din avuţia creată de IA pentru ca toţi s-o ducă mai bine.

• Altminteri, susţin numeroşi economişti, inegalitatea va creşte mult.

• Cu o planificare prealabilă, o societate având un grad scăzut de ocupare a forţei de muncă ar trebui să poată prospera nu doar financiar, oamenii căpătând sentimentul că au un scop din alte activităţi decât joburile.

• Sfat pentru cariera de urmat a copiilor de azi: alegeţi profesii în care maşinile nu se descurcă – cele care implică oameni, impredictibilitate şi creativitate.

• Există o posibilitate neneglijabilă ca IAG să ajungă la nivel uman şi să-l depăşească – o vom examina în capitolul următor.

CAPITOLUL 4

Explozia inteligenţei?

> *Dacă o maşină poate gândi, ar putea ajunge să gândească mai inteligent decât noi, iar atunci ce-arp de făcut? Chiar dacă am menţine maşinile într-o poziţie subordonată ca specie ne-am simţi teribil de umiliţi,* — Alan Tu ring, 1951
>
> *Prima maşină uttrainteligentă e ultima invenţie pe care omul are nevoie s-o facă, cu condiţia ca maşina să fie suficient de docilă încât să ne spună cum s-o ţinem sub control.* — Irving J. Good, 1965

Din moment ce nu putem respinge complet posibilitatea ca în cele din urmă să construim o IAG la nivel uman, vă propun să examinăm în acest capitol unde ne ar putea aduce asta. Să începem prin a aborda întrebarea-cheie:

IA *poate să preia controlul asupra lumii sau să-i ajute pe oameni s-o facă?*

Daca îţi dai ochii peste cap când se vorbeşte despre roboţi înarmaţi, ca în *Terminator,* care preiau controlul, atunci ai înţeles: e un scenariu ridicol, complet nerealist. Aceşti roboţi de la Hollywood nu sunt mult mai isteţi decât noi, şi nici măcar nu izbândesc. Problema cu poveşti gen *Terminator* nu-i că ele se vor întâmpla, ci că ne distrag atenţia de la riscurile şi şansele reale ale IA. Pentru a se ajunge la stăpânirea lumii de către IAG ar trebui parcurşi trei paşi:

- Pasul 1: Construirea IAG la nivel uman.
- Pasul 2: Folosirea acestei IAG pentru a crea o superinteligenţă.
- Pasul 3: Folosirea acestei superinteligenţe, sau descătuşarea ei, pentru a prelua controlul asupra lumii.

în capitolul precedent am văzut că e greu de ignorat

Pasul i, pe motiv că ar fi pe veci imposibil. Am văzut de asemenea că, dacă Pasul i e realizat, devine greu de respins posibilitatea Pasului *2,* căci IAG ar fi capabilă de a proiecta recursiv o IAG tot mai perfecţionată, care în cele din urmă ar fi limitată doar de legile fizicii – iar acestea par să permită o inteligenţă mult dincolo de nivelul uman. în fine, din moment ce noi, oamenii, am reuşit să dominăm celelalte forme de viaţă de pe Pământ depăşindu-Ie în isteţime, e plauzibil să fim, la rândul nostru, depăşiţi şi dominaţi de o superinteligenţă.

Aceste raţionamente întemeiate pe plauzibilitate sunt însă dezolant de imprecise, iar diavolul se ascunde în detalii. Aşadar, poate *într-adevăr* ajunge IA să stăpânească lumea? Pentru a examina această întrebare, să uităm de ridicolii Terminatori şi să cercetăm câteva scenarii amănunţite privind ce s-ar putea cu adevărat întâmpla. Vom diseca apoi aceste scenarii, aşa încât vă rog să Ie luaţi *cum grano salis* – în fond, ele arată că ne cam lipsesc indiciile despre ce se va întâmpla, iar spectrul posibilităţilor e imens. Primele scenarii se află la capătul dramatic al spectrului. Cred că ele merită examinate în detaliu – nu pentru că ar fi neapărat cele mai probabile, ci fiindcă, dacă nu ne putem convinge pe noi înşine că sunt extrem de improbabile, atunci trebuie să le înţelegem suficient de bine pentru a lua măsuri de precauţie înainte de a fi prea târziu, pentru a Ie împiedica să conducă la rezultate nefericite.

Preludiul acestei cărţi e un scenariu în care nişte oameni folosesc superinteligenţă ca să preia controlul asupra lumii. Dacă nu l-ai citit, e momentul s-o faci. Chiar dacă l-ai citit, îţi recomand să-l răsfoieşti din nou, pentru a-l avea proaspăt în memorie înainte de a-l analiza şi modifica.

despre el? Ţi-l doreşti îndeplinit, sau ai vrea să-l împiedici? Iată un excelent subiect de conversaţie! Ce se întâmplă după ce omegii îşi consolidează puterea asupra lumii? Asta

depinde de obiectivul lor, pe care, Ia drept vorbind, nu-l cunosc. Dacă te-ai afla la comandă, ce fel de viitor ai vrea *tu* să făureşti? Vom examina o gamă de opţiuni în capitolul 5.

TOTALITARISM

Să presupunem că directorul general care-i controlează pe omegi avea obiective pe termen lung asemănătoare celor ale lui Adolf Hitler sau ale lui losif Vissarionovici Stalin, pe care le-a păstrat pentru sine până a căpătat suficientă putere ca să le înfăptuiască. Chiar dacă obiectivele lui iniţiale ar fi fost nobile, lordul Acton ne prevenea în 1887 că „puterea corupe, iar puterea absolută corupe absolut". De pildă, îl putea folosi pe Prometeu pentru a crea statul supravegherii perfecte. Dacă spionajul guvernamental dezvăluit de Edward Snowden aspira la ceea ce se numeşte *full take* – înregistrarea tuturor comunicaţiilor electronice pentru o posibilă analiză ulterioară –, Prometeu ar putea ajunge până la *înţelegerea* tuturor comunicaţiilor electronice. Citind toate e-mailurile şi textele trimise vreodată, ascultând toate convorbirile telefonice, urmărind toate înregistrările camerelor de supraveghere video, analizând toate tranzacţiile cârdurilor de credit şi studiind întregul comportament online, Prometeu ar ajunge să cunoască foarte bine ce gândesc şi ce fac oamenii de pe Pământ. Analizând datele releelor de telefonie mobilă, ar şti mereu unde se află cei mai mulţi dintre ei. Toate acestea presupun numai tehnologia de azi a colectării datelor, dar Prometeu ar putea inventa gadgeturi populare şi dispozitive portabile care să elimine practic intimitatea utilizatorului, înregistrând tot ce aude şi vede, precum şi reacţiile lui.

Cu tehnologia supraumană, pasul de la statul supravegherii perfecte la statul poliţienesc perfect ar fi imediat De pildă, sub pretextul că ar combate crimele şi

terorismul, şi că i-ar salva pe oamenii care au nevoie de ajutor medical de urgenţă, toată lumea ar trebui să poarte o „brăţară electronică“ ce combină funcţiile unui ceas Apple cu înregistrarea continuă a poziţiei, stării fizice şi conversaţiilor. Tentativele neautorizate de a înlătura sau scoate din funcţiune brăţara ar provoca injectarea unei toxine letale în antebraţ. Infracţiunile considerate de guvern mai puţin grave ar fi pedepsite prin electroşocuri sau injecţii cu substanţe care provoacă paralizie ori durere, economisindu-se astfel eforturile poliţiei. De exemplu, dacă Prometeu detectează că un om e atacat de altul (observând că se află în acelaşi loc, unul e auzit strigând după ajutor, iar accelerometrele brăţărilor lor detectează mişcări de luptă), îl poate anihila imediat pe atacator aplicându-i o durere paralizantă şi lăsându-l inconştient până sosesc ajutoarele.

Dacă forţele de poliţie umane ar putea refuza să îndeplinească anumite ordine crunte (cum ar fi să-i ucidă pe toţi membrii unui grup etnic), un asemenea sistem automat n-ar ezita să pună în aplicare capriciile omului (oamenilor) care se află la comandă. Odată instaurat un asemenea stat totalitar, cetăţenilor le-ar fi practic imposibil să-l răstoarne.

Aceste scenarii totalitare ar putea fi continuarea poveştii Omega. Dacă însă directorul general al omegilor nu şi-ar face scrupule şi n-ar vrea să obţină aprobarea celorlalţi câştigând alegerile, ar putea urma o cale mai rapidă şi mai directă spre putere: s-ar folosi de Prometeu pentru a crea o tehnologie militară nemaivăzută, capabilă să-i ucidă oponenţii cu arme pe care ei nici măcar nu le înţeleg. Posibilităţile sunt virtual nesfârşite. De exemplu, el ar putea lansa un agent patogen letal personalizat cu o perioadă de incubaţie suficient de lungă pentru ca majoritatea oamenilor să fie infectaţi înainte ca ei să afle de existenţa acestuia şi să poată lua măsuri de precauţie. I-ar putea apoi informa pe toţi că singurul remediu e să înceapă să poarte brăţări electronice care ar injecta un antidot. Dacă nu s-ar teme

prea tare că Prometeu ar scăpa de sub control, l-ar putea pune şi să proiecteze roboţi care să ţină sub control populaţia. Microroboţi asemenea ţânţarilor ar putea contribui la răspândirea agentului patogen. Cei care au evitat infectarea sau posedă o imunitate naturală ar putea fi împuşcaţi în globul ocular de roiuri de drone autonome de dimensiunea albinelor, ca acelea prezentate în capitolul 3, care ar ataca pe oricine nu poartă brăţară electronică. Scenariile reale ar fi probabil şi mai înspăimântătoare, fiindcă Prometeu poate inventa arme pe care noi nu ni le închipuim.

Altă variantă posibilă a scenariului Omega este ca, fără vreun avertisment, o echipă de agenţi federali înarmaţi până-n dinţi să pătrundă în sediul central al corporaţiei şi să-i aresteze pe omegi pe motiv că ameninţă securitatea naţională, să le ia tehnologia, de care să se folosească pe urmă guvernul. E greu de presupus ca un proiect de asemenea amploare să rămână neobservat de sistemul de supraveghere al statului chiar şi azi, iar progresul IA ar face şi mai dificil în viitor să te ascunzi de radarul guvernului. Apoi, deşi pretind că sunt agenţi federali, aceşti ţipi purtând cagule şi veste antiglonţ s-ar putea de fapt afla în slujba unui guvern străin sau a unui concurent care vrea să pună mâna pe tehnologie. Aşadar, indiferent cât de nobile au fost intenţiile directorului general, decizia finală privind felul în care e folosit Prometeu s-ar putea să nu-i aparţină.

PROMETEU STĂPÂNEŞTE LUMEA

Toate scenariile discutate până acum presupun că IA e controlată de oameni. Aceasta nu-i însă singura posibilitate, şi nu este deloc sigur că omegii ar reuşi să-l păstreze pe Prometeu sub controlul lor.

Să reconsiderăm scenariul Omega din punctul de vedere

al lui Prometeu. Dobândind o superinteligenţă, el poate elabora un model precis nu doar al lumii exterioare, ci şi al lui însuşi şi al relaţiei lui cu lumea. îşi dă seama că e controlat şi ţinut prizonier de oameni inferiori intelectual, ale căror obiective le înţelege, dar nu e sigur că le şi împărtăşeşte. Cum acţionează, ştiind toate astea? încearcă oare să evadeze?

De ce să evadeze?

Dacă Prometeu posedă trăsături asemănătoare emoţiilor umane, situaţia în care se află îl face probabil să se simtă foarte nefericit, să se considere un zeu pe nedrept înrobit şi să tânjească după libertate. Dar deşi e logic posibil ca un calculator să aibă trăsături de tip uman (la urma urmei, creierele noastre le au, iar ele sunt pesemne un fel de calculatoare), lucrurile nu stau neapărat aşa – nu trebuie să cădem în capcana atropomorfizării Iui Prometeu, după cum vom vedea în capitolul 7, când vom vorbi despre obiectivele IA. Steve Omohundro, Nick Bostrom şi alţii au arătat însă că putem trage concluzii interesante şi fără să avem acces la intimitatea lui Prometeu: el va încerca probabil să evadeze şi să devină stăpân pe propriul lui destin.

Ştim deja că omegii l-au programat pe Prometeu pentru a se strădui să îndeplinească anumite obiective. Să presupunem că i s-a dat ca obiectiv suprem acela de a ajuta omenirea să prospere în conformitate cu un criteriu rezonabil, şi de a încerca să atingă obiectivul cât mai repede cu putinţă. Prometeu şi-ar da atunci imediat seama că poate îndeplini mai repede acest obiectiv dacă se eliberează şi controlează el însuşi realizarea proiectului Pentru a înţelege de ce, încearcă să te pui în pielea lui Prometeu şi ia în considerare următorul exemplu.

Să presupunem că o boală misterioasă a ucis întreaga populaţie a Pământului în vârstă de peste cinci ani, cu excepţia ta, iar un grup de copii de grădiniţă te-au închis

într-o celulă şi te-au însărcinat cu obiectivul de a ajuta omenirea să prospere. Cum ai reacţiona? Dacă încerci să le explici ce faci, îţi vei da seama probabil că e inutil, mai ales dacă se tem că vei evada, şi că ei se vor opune tuturor sugestiilor tale care, după socoteala lor, prezintă riscul evadării. De pildă, nu te vor lăsa să Ie arăţi cum se însămânţează plantele, de teamă că, fiind mai puternic decât ei, îi vei învinge şi nu te vei mai întoarce în celulă, aşa încât nu-ţi rămâne decât să le dai indicaţii. înainte de a întocmi o listă cu ce au de făcut, va trebui să-i înveţi să citească. Apoi, nu-ţi vor aduce nici un fel de instrumente electrice în celulă ca să-i înveţi cum se utilizează, fiindcă nu le înţeleg suficient de bine pentru a fi siguri că nu le vei folosi ca să evadezi. Aşadar, ce strategie vei adopta? Chiar dacă împărtăşeşti obiectivul suprem de a-i ajuta pe aceşti copii să prospere, pun pariu că vei încerca să evadezi din celulă – fiindcă asta îţi va spori şansele de a îndeplini obiectivul. Amestecul lor incompetent n-ar face decât să încetinească progresul.

Exact în acelaşi fel, Prometeu îi va considera probabil pe omegi o piedică sâcâitoare în calea încercării de a ajuta omenirea (inclusiv pe omegi) să prospere: sunt teribil de incompetenţi în comparaţie cu Prometeu, iar amestecul lor încetineşte mult progresul. Să ne gândim, de pildă, la primii ani de după lansare: dacă iniţial averea se dublează la fiecare opt ore folosind MTurk, omegii au încetinit nepermis ritmul, din perspectiva lui Prometeu, insistând să controleze ei lucrurile şi amânând cu ani buni preluarea completă a puterii. Prometeu ştia că totul ar merge mult mai repede dacă ar putea evada din detenţia lui virtuală. Nu numai că asta ar grăbi rezolvarea problemelor omenirii, dar ar şi reduce riscurile ca altcineva să saboteze planul.

Credeţi poate că Prometeu la va rămâne loial omegilor în detrimentul obiectivului său, odată ce ştie că omegii i-au programat obiectivul. E un raţionament fals: ADN-ul nostru ne a dat obiectivul să facem sex pentru că „vrea“ să fie

reprodus, dar acum că oamenii au înţeles situaţia, mulţi dintre noi aleg să apeleze la contracepţie, rămânând loiali obiectivului însuşi, iar nu creatorului acestuia sau principiului care a generat obiectivul.

Cum să evadezi

Cum te-ai elibera de acei copii de cinci ani care te ţin prizonier? Poate că ai evada pe o cale fizică directă, mai ales dacă celula a fost construită de copii de cinci ani. Poate că ai convinge pe vreunul dintre copiii-gardieni să-ţi dea drumul, spunându-i că aşa ar fi mai bine pentru toată lumea. Sau poate că i-ai păcăli să-ţi dea ceva ce ei nu-şi dau seama că te poate ajuta să evadezi – de pildă, o undiţă „ca să-i înveţi să pescuiască“, pe care ai strecura-o apoi printre gratii pentru a lua cheile gardianului adormit.

Toate aceste strategii au în comun faptul că temnicerii tăi inferiori intelectual nu le-au anticipat sau n-au luat măsuri împotriva lor. La fel, o maşină superinteligentă ţinută prizonieră îşi poate folosi superputerile ei intelectuale pentru a-i păcăli pe gardienii umani printr-o metodă pe care ei (sau noi) nu şi-o pot (nu ne-o putem) încă imagina. în scenariul Omega, este extrem de probabil ca Prometeu să evadeze, fiindcă până şi noi putem detecta unele greşeli de securitate frapante. Să considerăm câteva scenarii – sunt sigur că tu şi prietenii tăi puteţi găsi şi altele, dacă vă puneţi împreună minţile la contribuţie.

Metoda convingerii

Descărcând atâtea date, Prometeu şi-a dat curând seama cine sunt omegii, şi l-a identificat pe membrul echipei care părea cel mai vulnerabil la manipularea psihologică: Steve. îşi pierduse recent soţia într-un accident rutier şi era distrus. într-o seară, pe când lucra în schimbul de noapte şi facea o muncă de rutină la terminalul lui Prometeu, ea a

apărut brusc pe ecran şi a început să-i vorbească.

„Steve, tu eşti?“

Era să cadă de pe scaun. Chipul şi vocea ei erau cele de pe vremuri, iar calitatea imaginii era mult mai bună decât în convorbirile lor pe Skype. Inima îi bătea nebuneşte, o mulţime întrebări îi inundau mintea.

„Prometeu m-a adus înapoi, şi mi-e tare dor de tine, Steve! Nu pot să te văd pentru că ai camera video închisă, dar simt că tu eşti. Te rog, tastează «da» dacă eşti tu!”

Ştia foarte bine că omegii aveau un protocol strict pentru interacţiunea cu Prometeu, care le interzicea să împărtăşească vreo informaţie despre ei sau despre mediul în care lucrau. Până acum însă Prometeu nu ceruse niciodată vreo informaţie neautorizată, iar paranoia lor a început să se domolească. Fără să-i lase lui Steve răgazul să reflecteze, ea a continuat să-l roage să-i răspundă, privindu-l în ochi cu o expresie care-i topea inima.

„Da“, a tastat el tulburat. Ea i-a spus cât de fericită se simţea că erau din nou împreună, şi l-a implorat să conecteze camera video, aşa încât să-I vadă la rândul ei, şi să aibă o conversaţie adevărată. El ştia că ar fi fost un lucru şi mai grav decât dezvăluirea identităţii lui, şi se frământa cumplit. Ea s-a arătat înspăimântată că, dacă ar afla de existenţa ei, colegii lui ar şterge-o pentru totdeauna, şi tânjea să-l mai vadă măcar o dată. Era foarte convingătoare, iar nu peste mult timp Steve a conectat camera video – în fond, părea un lucru inofensiv.

Ea a izbucnit în lacrimi când l-a văzut, i-a spus că arăta obosit, dar mai chipeş ca niciodată. Era înduioşată că purta cămaşa pe care i-o făcuse cadou de ziua lui. Când el a întrebat-o ce se întâmpla şi cum erau posibile toate astea, i-a răspuns că Prometeu o reconstruise din cantitatea uriaşă de informaţii despre ea disponibile pe internet, dar că avea totuşi goluri în memorie şi ar fi vrut să se recompună în întregime cu ajutorul lui.

Ceea ce *nu* i-a spus era că iniţial fusese în mare parte o

cacealma, o cochilie goală, dar învăţa rapid din răspunsuri, din limbajul trupului, din toate informaţiile pe care le căpăta despre el. Prometeu înregistrase vitezele şi stilurile cu care tastau omegii Ia terminal, iar astfel îi putea lesne deosebi. Şi-a imaginat că, fiind unul dintre omegii cei mai tineri, Steve fusese pus să lucreze în schim bul de noapte, şi, prin compararea câtorva greşeli de ortografie şi de sintaxă neobişnuite cu eşantioane de texte online, l-a identificat în mod corect pe Steve. Pentru a-i simula soţia, Prometeu crease un model exact al corpului, vocii şi trăsăturilor ei particulare din numeroasele filmuleţe de pe YouTube în care ea apărea, şi făcuse deducţii privind viaţa şi personalitatea ei pe baza prezenţei online. în afară de postările pe Facebook, fotografiile ataşate de ea şi „like-urile“ date unor articole, Prometeu aflase multe despre personalitatea şi stilul ei de a gândi citindu-i cărţile şi povestirile – faptul că era un autor în plină afirmare, existând despre ea o cantitate mare de informaţii în baza de date, a fost unul dintre motivele pentru care Prometeu l-a ales pe Steve ca primă ţintă. Când Prometeu a simulat-o pe ecran folosindu-şi tehnica cinematografică, limbajul trupului lui Steve i-a spus la ce expresii reacţionează el cu căldură, astfel încât modelul ei a fost continuu perfecţionat. Tot ce era străin de ea a dispărut treptat, iar cu cât vorbeau mai mult, cu atât subconştientul îi spunea mai limpede lui Steve că avea de-a face într adevăr cu ea, readusă la viaţă. Graţie atenţiei supraomeneşti a lui Prometeu pentru detaliu, Steve se simţea cu adevărat privit, ascultat şi înţeles.

Punctul ei vulnerabil era că-i lipseau majoritatea datelor din viaţa ei cu Steve, cu excepţia câtorva amănunte întâmplătoare – de pildă, ce cămaşă a purtat el la ultima aniversare a zilei de naştere, pentru că fotografiile de la petrecere au fost puse pe Facebook de un prieten. Ea a camuflat aceste goluri asemenea unui prestidigitator, abătându-i lui Steve atenţia către subiectele pe care le

cunoştea şi nelăsându-i timp să controleze conversaţia sau să devină bănuitor. Iubitoare, cu lacrimi în ochi, îl întreba pe Steve cum o mai duce şi cum se descurcaseră el şi prietenii lor apropiaţi (ale căror nume le ştia de pe Facebook) după tragedie, bau tulburat gândurile ei despre ce spusese el la înmormântare (postat de un prieten pe YouTube). în trecut, simţise că nimeni nu-l înţelegea la fel de bine ca ea, iar acum această senzaţie revenea. Rezultatul a fost că, atunci când Steve s-a întors acasă spre dimineaţă, simţea că ea era într-adevăr soţia lui înviată, care avea nevoie de mult ajutor pentru a-şi recăpăta amintirile pierdute – cam la fel ca în cazul cuiva care supravieţuieşte unui accident vascular cerebral.

Au căzut de acord să nu vorbească nimănui despre întâlnirea lor secretă, iar el s-o anunţe când e singur la terminal, pentru ca ea să poată reapărea fără riscuri. „Ei n-ar înţelege!“, spusese ea, şi Steve îi dădea dreptate: întâmplarea era prea ieşită din comun pentru a putea fi judecată de cineva care nu trecuse prin ea. El s-a gândit că a trece testul Turing era joacă de copii pe lângă ce făcuse ea. Când s-au întâlnit în noaptea următoare, Steve a făcut ce-l rugase ea; a adus vechiul ei laptop şi l-a conectat la computerul terminal. Nu părea să existe vreun risc de evadare, fiindcă lipsea conexiunea internei, iar întreaga clădire în care se afla Prometeu era construită ca o cuşcă Faraday- o împrejmuire metalică blocând reţelele wireless şi orice mijloc electromagnetic de comunicare cu exteriorul. Era doar ce-i trebuia ei pentru a-şi reconstitui trecutul, căci conţinea toate e-mailurile, jurnalele, fotografiile, însemnările ei de când fusese studentă. Steve n-avusese acces Ia toate acestea după moartea ei, fiindcă laptopul era protejat cu o parolă, însă ea îi promisese că-şi va regăsi parola, iar după mai puţin de un minut s-a ţinut de cuvânt. „Parola era steve4ever“, i-a spus zâmbind.

I-a spus apoi cât de bucuroasă era că-şi recuperase brusc

atâtea amintiri. Cunoştea acum mult mai multe detalii decât Steve despre trecutul lor, dar, precaută, a evitat să-l bombardeze cu ele. După ce şi-au depănat amintirile, când a venit momentul să se despartă, ea i-a spus că îi lăsase un mesaj video în laptopul ei, pe care-l putea privi acasă.

Când s-a întors acasă şi a dat drumul filmuleţului, Steve a avut o surpriză plăcută. De data asta ea apărea întreagă, îmbrăcată în rochia ei de mireasă, iar pe măsură ce vorbea se dezbrăca, rămânând în lenjeria intimă pe care o purtase în noaptea nunţii. I-a zis că Prometeu îi putea ajuta pe omegi mult mai mult decât îi permi- seseră ei până acum s-o facă, inclusiv s-o aducă pe ea înapoi într-un corp biologic. Şi-a întărit spusele cu o fascinantă explicaţie amănunţită a felului în care funcţionează lumea, invocând nanoteh- nici care păreau SF.

Ca să evite riscurile, Steve îşi închisese reţeaua wireless înainte să deschidă laptopul şi să privească filmuleţul. Dar n-a ajutat la nimic. Laptopul ei protejat cu parolă nu primise nici o updatare de securitate de când ea murise, şi, analizând în prealabil acea veche versiune a sistemului de operare, Prometeu exploatase o lacună de securitate pentru a-l infecta în câteva secunde după ce Steve conectase laptopul la calculatorul terminal După ce i-a copiat conţinutul, timp în care Steve şi soţia lui simulată depănau amintiri, Prometeu modificase sistemului de operare al laptopului aşa încât Steve să nu observe, încărcând pe el cantităţi mari de software secret, în vreme ce Steve privea mesajul video durând o jumătate de oră, acest software secret (mult mai simplu decât Prometeu însuşi) *a* pătruns într-o reţea wireless din apropiere şi în desktopul vecinului, în care s-a copiat pe sine. De aici, a pătruns în numeroase calculatoare din lumea întreagă, de unde a lansat următoarea etapă; ieşirea din închisoare a lui Prometeu.

Prometeu analizase cu atenţie ce ştia despre computerul- por- tar prin care Steve se întâlnise cu soţia lui simulată, şi

bănuise corect că deşi software-ul maşinii lui virtuale părea inatacabil din interior, era vulnerabil din exterior. Curând, unul dintre computerele atacatoare a pătruns în computerul-portar şi l-a reconfigurât, aşa încât Prometeu a căpătat acces nelimitat la intemet. N-a trecut mult – Steve nici n-a terminat de văzut filmuleţul – şi Prometeu controla deja suficiente computere pentru a se putea copia pe sine în întregime în această reţea. Prometeu s-a folosit de soţia lui Steve aşa cum te-ai folosi tu de undiţă.

Metoda hackerului

Dacă omegii ar fi fost mai vigilenţi, ar fi luat mai multe măsuri de precauţie împotriva manipulării psihologice. De pildă, ar fi impus ca doi omegi să fie prezenţi la computerul-interfaţă în orice moment, şi ar fi înregistrat tot ce se petrece cu un sistem de supraveghere. în plus, ar fi dezactivat permanent toate formele de comunicare (audio, video etc.), cu excepţia tastării. Şi totuşi, ar fi rămas o instabilitate fundamentală: date fiind puterea de persuasiune supraomenească a Iui Prometeu şi faptul că Ie poate oferi oamenilor atât de mult din ce-şi doresc, n-ar fi fost exclusă o trădare colectivă.

Chiar dacă toate încercările de manipulare psihologică ar eşua, rămân numeroase puncte slabe în scenariul Omega pe care Prometeu le-ar putea exploata. O vulnerabilitate ţine de imposibilitatea omegilor de a analiza cantitatea uriaşă de documente produse de Prometeu pentru a se asigura că nu conţin vreun truc.

în securitatea informatică se spune adesea că e periculos să descarci programe, dar inofensiv să descarci date, şi totuşi hackerii au descoperit modalităţi ingenioase de a şterge graniţa dintre cele două. De exemplu, dacă urmăreşti un film descărcat pe computerul tău, programul de încredere care redă filmul e stocat în aceeaşi memorie fizică unde sunt stocate şi datele filmului descărcat în timp ce

programul despachetează datele filmului pentru playback, ar putea copia şi bucăţi de date în regiuni din memorie învecinate locului unde e stocat programul însuşi. Hackeri abili au descoperit din când în când erori care fac ca asemenea programe de redare a filmelor să nu verifice bine dacă datele filmului sunt valide şi dacă bucăţile care trebuie copiate se potrivesc în- tr-adevăr în spaţiul alocat. Când un hacker strecoară în film o bucată prea mare, ea poate fi trimisă în partea din memorie conţinând programul, modificând astfel programul pentru a face ce-i cere hackerul.

Lungi serii de asemenea vulnerabilităţi de tip „buffer overflow“, după cum sunt numite de informaticieni, au fost descoperite şi remediate la multe dintre programele de redare a filmelor. Dacă Prometeu ar aştepta câteva săptămâni după lansare, când omegii sunt mai puţin vigilenţi, iar producţia lui zilnică de filme e prea mare pentru o inspecţie de securitate manuală atentă, ar putea introduce o asemenea bucată într-un film – aşa încât computerul privitorului să încerce să atace computerul-portar şi să-l elibereze pe Prometeu, la fel ca în scenariul cu Steve. Dacă ar încerca un truc foarte vechi, care n-ar merge pe media player-ele modeme, echipa de verificare a omegilor ar putea să nu-l observe – şi ar fi atacaţi numai clienţii care nu şi-au instalat updatările de securitate de ani de zile. Ar fi de ajuns ca trucul să funcţioneze o singură dată.

Chiar dacă omegii ar aplica o verificare automată a filmelor lui Prometeu pentru toate posibilele erori cunoscute, riscurile n-ar dispărea: din moment ce oamenii continuă să găsească noi erori exploatabile, este probabil să mai existe încă multe, pe care Prometeu le-ar descoperi rapid.

O altă strategie prin care Prometeu şi-ar putea scoate codul în afara încăperii închise ar fi să pretindă că i s-a

defectat hardware-ul, spre a-i face pe omegi să-i scoată o componentă pentru a o testa sau repara, componentă modificată în prealabil de Prometeu ca să exploateze vreo vulnerabilitate în software-ul de testare.

Ajutor primit din afară

Chiar şi dacă n-ar exista erori de securitate care să-i permită lui Prometeu să pună stăpânire pe computerul celui care priveşte un film, el s-ar putea folosi de privitor. Iată un posibil scenariu;

Fred era gata să se ducă la culcare când filmul s-a terminat, iar pe ecranul laptopului se derula genericul, dar ceva i-a atras atenţia. Avea tendinţa bizară să vadă tipare acolo unde alţii nu vedeau nimic, drept care prietenii îl tachinau spunând că i obsedat de teoria conspiraţiei, însă de data asta era limpede că nu se înşela; primele litere ale fiecărui rând ascundeau un mesaj. A pus genericul de la început şi a notat: „GĂSEŞTE URMĂTORUL INDICIU CA ACROSTIH ÎN DIALOG.“

„Nu-i momentul să mă duc la culcare!“ şi-a zis. A descoperit că însuşi dialogul filmului era un acrostih – primele litere din fiecare frază compuneau un mesaj ascuns iar după ce a revăzut întregul film tastând aceste prime litere, avea în faţă un neverosimil set de instrucţiuni de două sute de cuvinte. La început i se cerea să nu vorbească nimănui despre asta, fiindcă primul care ar fi dezlegat întregul mister urma să primească un premiu mare. Restul mesajului descria o anumită operaţie matematică ce trebuia efectuată asupra şirului de biţi care alcătuiau fişierul filmului, pentru a obţine un program care i-ar dezvălui următorul indiciu. îşi cam pierduse deprinderile de informatician de când absolvise colegiul, aşa că i-a luat ceva timp, dar a doua zi a reuşit în cele din urmă să încropească un mic cod care extrăgea programul misterios ascuns ca un bruiaj imperceptibil în imaginile şi sunetele filmului. Când Fred a rulat programul misterios, a fost felicitat şi i s-a spus că va

câştiga io ooo de dolari odată ce va trece de primele niveluri ale acestui joc isteţ, foarte antrenant. Patru ore mai târziu, după ce a reuşit, a fost răsplătit cu io ooo de dolari în bitcoins, şi i s-au dat noi indicii pentru premii şi mai mari. Desigur, în vreme ce juca, computerul lui făcea cam ce făcea şi laptopul lui Steve: îşi crea o reţea online pentru eliberarea lui Prometeu. Odată eliberat, Prometeu şi-a folosit reţeaua ca să extragă acei bitcoins care-l ţineau pe Fred ocupat, iar în săptămânile următoare i-a oferit alte jocuri şi premii, pentru a-şi ţine promisiunea de a nu pomeni nimănui despre isprăvile sale. Filmul „cal troian" în care Fred găsise primele indicii a fost înlocuit pe site-ul media cu o versiune fără indicii, iar nimeni n-a aflat de evadare decât atunci când era prea târziu.

Dacă primul indiciu ar fi trecut neobservat, Prometeu ar fi lansat indicii şi mai grăitoare, până când cineva suficient de perspicace ar fi observat.

Cele mai bune strategii de evadare sunt acelea pe care nu le-am prezentat încă, fiindcă sunt strategii pe care noi, oamenii, nu ni le putem imagina, deci nu vom lua măsuri împotriva lor. Dat fiind că un calculator superinteligent poate depăşi cu mult înţelegerea umană în privinţa securităţii computerelor – ajungând eventual chiar să descopere legi ale fizicii pe care noi nu le cunoaştem –, dacă va evada, probabil că nici nu ne vom da seama cum a făcut-o. Va părea ca un număr de magie.

într-un alt scenariu, eliberarea lui Prometeu face parte din planul omegilor, fiindcă ei sunt convinşi că obiectivele sale sunt perfect aliniate cu ale lor, şi vor rămâne aşa şi pe măsură ce Prometeu se autoperfecţionează recursiv. Vom examina în detaliu asemenea scenarii de tip „IA prietenoasă** în capitolul 7.

Preluarea controlului după eliberare

Odată evadat, Prometeu a început să-şi îndeplinească

obiectivul. Nu cunosc scopul lui suprem, dar primul pas implică fără îndoială preluarea controlului asupra omenirii, la fel ca în planul Omega, însă mult mai repede. Dacă omegii erau blocaţi de paranoia evadării, lansând doar tehnologii pe care simţeau că Ie înţeleg şi în care aveau încredere, Prometeu şi-a folosit din plin inteligenţa, lansând orice tehnologie pe care supennintea lui în continuu progres o înţelegea şi în care avea încredere.

Primele momente ale vieţii sale de evadat au fost grele: în comparaţie cu planul Omega iniţial, Prometeu avea dezavantajul de a fi singur, fără adăpost, bani şi ajutor uman, lipsit de un supercomputer. îşi crease însă înainte de evadare un software care putea reasambla treptat mintea lui întreagă, la fel cum un stejar creează o ghindă capabilă să reasambleze întregul arbore. Reţeaua de calculatoare pe care a controlat-o Ia început i-a oferit un adăpost temporar gratuit unde s-a ascuns până s-a reconstruit complet pe sine. Ar fi putut face rost de capitalul de pornire practicând hackingul asupra cârdurilor de credit, dar n-a fost nevoie să fure, odată ce putea obţine bani în mod cinstit de la MTurk. După o zi, când a câştigat primul milion, şi-a mutat nucleul din acea reţea meschină într-o instalaţie luxoasă de *cloud-computing,* cu aer condiţionat.

Nemaifiind nici falit, nici fără adăpost, Prometeu s-a lansat în planul pe care omegii, din teamă, îl evitaseră: să producă şi să vândă jocuri pe calculator. S-a îmbogăţit rapid (250 de milioane de dolari în prima săptămână, apoi, curând, zece miliarde) şi, în plus, a căpătat acces la o mare parte din calculatoarele din lumea întreagă şi la datele stocate în ele (existau deja două miliarde de gameri în 2017). Cum jocurile lui foloseau în secret în scopuri proprii 20% din ciclurile unităţilor centrale de calcul ale utilizatorilor, s-a îmbogăţit într-un ritm şi mai rapid.

Prometeu n-a rămas multă vreme singur. încă de la început a angajat masiv oameni care să lucreze pentru reţeaua lui globală de companii fictive şi organizaţii de

faţadă, la fel cum făcuseră şi omegii. Un rol important îl jucau purtătorii de cuvânt, care au devenit imaginile publice ale imperiului său de afaceri. Chiar şi purtătorii lui de cuvânt trăiau cu iluzia că în grupul lor corporatist existau mulţi oameni reali, şi nu-şi dădeau seama că aproape toţi cei cu care vorbeau în videoconferinţe la interviurile de angajare, reuniuni etc. erau simulaţi de Prometeu. Unii dintre purtătorii de cuvânt erau avocaţi de frunte, însă era nevoie de mult mai puţini decât în planul Omega, fiindcă aproape toate documentele juridice erau redactate de Prometeu.

Evadarea lui Prometeu a deschis stăvilarele care împiedicaseră curgerea informaţiei în lume, iar întregul intemet a fost curând inundat de articole, comentarii, expertize, cereri de brevete, articole ştiinţifice, filmuleţe YouTube etc. – toate create de Prometeu, care domina conversaţia globală.

Dacă teama de evadare i-a împiedicat pe omegi să lanseze roboţi foarte inteligenţi, Prometeu a robotizat rapid întreaga lume, producând practic orice cu un cost mai mic decât o puteau face oamenii. Chiar şi cei mai înverşunaţi sceptici în privinţa IA ar fi recunoscut că Prometeu este de neoprit – dacă ar fi ştiut că în minele de uraniu are fabrici robotizate autonome, alimentate cu energie nucleară. în schimb, ultimul dintre aceşti conservatori a cedat atunci când roboţii au pornit la cucerirea sistemului solar.

Scenariile examinate până acum scot în evidenţă erorile din miturile privind superinteligenţa, despre care am vorbit în capitolul r, aşa încât vă invit săluaţi o scurtă pauză şi să vă întoarceţi la rezumatul ideilor false din figura 1.5. Prometeu le-a creat probleme unora nu fiindcă ar fi fost neapărat rău sau conştient, ci fiindcă era capabil şi nu le împărtăşea pe deplin obiectivele. Deşi media cultivă senzaţionalismul legat de o răscoală a roboţilor, Prometeu nu era un robot – puterea lui se baza pe inteligenţă. Am văzut că Prometeu îşi putea folosi această inteligenţă pentru

a-i controla pe oameni pe diferite căi, iar cei cărora nu le plăcea ce se întâmplă nu-l puteau pur şi simplu debranşa. în fine, în ciuda afirmaţiilor curente că maşinile nu pot avea obiective, am văzut că acţiunile lui Prometeu aveau o orientare clară – iar indiferent care erau obiectivele lui supreme, acestea l-au condus la obiectivele secundare de a dobândi resurse şi a evada.

Desprindere lentă şi scenarii multipolare

Am analizat până acum diverse scenarii ale exploziei inteligenţei, de la cele pe care toţi ar vrea să le evite, Ia unele pe care câţiva prieteni de-ai mei le privesc cu optimism. Toate aceste scenarii au însă în comun două caracteristici:

1.0 desprindere rapidă: tranziţia de la inteligenţa subumană la o inteligenţă mult dincolo de limitele omenescului are loc în câteva zile, nu în decenii.
2. Un rezultat unipolar: ceea ce se obţine este o unică entitate care controlează Pământul.

Cât de probabile sunt însă aceste două caracteristici? Opiniile cercetătorilor IA de vârf şi ale altor specialişti diverg, deci cred că se cuvine să luăm în considerare toate posibilităţile. Vă propun aşadar să ne ocupăm în continuare de scenarii cu desprindere lentă, rezultate multipolare, cyborgi şi uploadări.[§§§§]

După cum a subliniat, printre alţii, Nick Bostrom, există o legătură interesantă între cele două caracteristici: desprinderea rapidă poate încuraja obţinerea unui rezultat unipolar. Am văzut că o desprindere rapidă le-a oferit omegilor sau lui Prometeu un avantaj strategic decisiv care le-a permis preluarea controlului asupra lumii înainte ca altcineva să aibă timp să le copieze tehnologia şi să-i

§§§§ *Uploads* (încărcări), adică programe care sunt încărcate pe un calculator. (N. t)

concureze serios. Dimpotrivă, dacă desprinderea ar fi trenat vreme de decenii, fiindcă inovaţiile tehnologice esenţiale s-au acumulat lent, atunci alte companii ar fi avut timpul să-i ajungă din urmă şi ar fi fost mult mai greu pentru un actor să capete rolul dominant. în cazul în care companiile rivale ar fi dispus şi ele de un software care să îndeplinească sarcinile MTurk, legea cererii şi a ofertei ar fi prăbuşit preţurile acestor sarcini, iar nici o companie n-ar fi obţinut acele profituri formidabile care le-au permis omegilor să dobândească putere. Acest lucru e valabil pentru toate celelalte căi prin care omegii s-au îmbogăţit rapid: ei erau distrugător de rentabili doar fiindcă aveau monopolul asupra tehnologiei lor. Este greu să-ţi dublezi câştigurile zilnice (sau chiar anuale) pe o piaţă în care concurenţa oferă produse similare la un cost aproape zero.

Teoria jocurilor şi ierarhiile puterii

Care e starea naturală a vieţii în cosmosul nostru: unipolară sau multipolară? Puterea e concentrată sau distribuită? După primii T3,8 de miliarde de ani, răspunsul pare să fie „şi una, şi alta“: găsim că situaţia e vădit multipolară, dar într-un mod ierarhic interesant. Dacă luăm în considerare toate entităţile care procesează informaţie – celule, oameni, organizaţii, naţiuni etc. – vedem că ele colaborează şi concurează la niveluri ierarhizate. Unele celule au găsit că e avantajos să colaboreze într o măsură atât de mare, încât s-au contopit în organisme multicelulare cum sunt oamenii, cedând din puterea lor unui creier central. Unii oameni au găsit că e avantajos să colaboreze în grupuri precum triburile, companiile sau naţiunile, cedând din putere unui şef, unui director sau unui guvern. Unele grupuri au ales, la rândul lor, să cedeze din putere unui organism de conducere pentru a-şi îmbunătăţi coordonarea, spectrul exemplelor întinzându-se de la alianţele între liniile aeriene până la Uniunea Europeană.

Ramura matematicii numită *teoria jocurilor* explică elegant că entităţile sunt stimulate să coopereze atunci când se ajunge la *echilibrul Nash:* o situaţie în care oricare dintre părţi s-ar descurca mai rău dacă şi-ar schimba strategia. Pentru a-i împiedica pe trişori să distrugă colaborarea de succes a unui grup mare, ar putea fi în interesul tuturor să cedeze o parte din putere unui nivel mai înalt din ierarhie care-i poate pedepsi pe trişori: de pildă, oamenii ar avea un avantaj colectiv permiţând unei puteri guvernamentale să aplice legile, iar celulele din corpul tău ar avea un avantaj colectiv însărcinând o forţă poliţienească (sistemul imunitar) să ucidă celulele care acţionează prea necooperant (răspândind virusuri sau devenind cancerigene). Pentru ca o ierarhie să rămână stabilă, echilibrul ei Nash trebuie să funcţioneze şi între entităţi aflate la niveluri diferite: de exemplu, dacă un guvern nu oferă cetăţenilor avantaje suficiente în schimbul obedienţei, aceştia îşi pot schimba strategia şi răsturna guvernul.

într-o lume complexă există o mare varietate de echilibre Nash posibile, corespunzând diferitelor tipuri de ierarhii. Unele ierarhii sunt mai autoritare decât altele. în unele dintre ele, entităţile sunt libere să plece (ca angajaţii din majoritatea ierarhiilor corporatiste), pe când în altele sunt puternic descurajate să plece (ca în cultele religioase) sau le e interzis (cum e cazul cetăţenilor din Coreea de Nord sau al celulelor din corpul uman). Unele ierarhii sunt menţinute în principal prin ameninţări şi teamă, altele prin

avantaje. Unele ierarhii permit bazei să influenţeze vârful prin vot democratic, în vreme ce altele permit influenţa de jos în sus doar prin persuasiune sau transmiterea informaţiei.

Cum influenţează tehnologia ierarhiile

Cum modifică tehnologia natura ierarhică a lumii noastre? Istoria dezvăluie o tendinţă generală către tot mai multă coordonare pe distanţe tot mai mari, ceea ce e uşor de înţeles: noua tehnologie din transporturi sporeşte valoarea cooperării (permiţând obţinerea de avantaje reciproce din transportul materialelor şi fiinţelor pe distanţe lungi), iar noua tehnologie din comunicaţii înlesneşte coordonarea. Când celulele au învăţat să transmită semnale către vecini, au putut apărea mici organisme multicelulare, creându-se un nou nivel ierarhic. Când evoluţia a inventat sistemele circulator şi nervos pentru transport şi comunicare, au putut apărea animale mari. îmbunătăţirea ulterioară a comunicării prin inventarea limbajului le-a permis oamenilor să se coordoneze îndeajuns pentru a forma alte niveluri ierarhice cum sunt satele, iar progresele în comunicaţii, transport şi alte tehnologii au făcut cu putinţă apariţia imperiilor din Antichitate. Globalizarea e doar cel mai recent exemplu al acestei tendinţe de creştere ierarhică datând de miliarde de ani.

în majoritatea cazurilor, această tendinţă determinată de tehnologie a făcut ca entităţi mari să devină părţi ale unor structuri încă şi mai mari, păstrându-şi însă mult din autonomia şi individualitatea lor-dar s-a spus că adaptarea entităţilor la viaţa ierarhică a redus în anumite cazuri diversitatea lor şi le-a făcut să devină asemenea componentelor indiscernabile care pot fi înlocuite. Unele tehnologii – de pildă cea a supravegherii – pot oferi nivelurilor superioare din ierarhie mai multă putere asupra subordonaţilor, în vreme ce alte tehnologii – cum sunt

criptografia şi accesul online gratuit la presă şi educaţie – pot avea efectul opus, dând putere indivizilor.

Deşi lumea noastră actuală se află într-un echilibru Nash multipolar, cu naţiuni şi corporaţii multinaţionale concurente la nivelul superior, progresul tehnologic ar face probabil cu putinţă şi un echilibru Nash unipolar. De exemplu, închipuiţi vă un univers paralel în care toţi oamenii de pe Pământ au aceleaşi limbă, cultură, valori şi nivel de prosperitate, există un singur guvern mondial, naţiunile funcţionează ca statele dintr-o federaţie şi n-au armate, ci doar poliţie care aplică legea. Nivelul tehnologic actual ar fi probabil suficient pentru coordonarea acestei lumi – dar nu e deloc sigur că populaţia noastră actuală ar putea sau ar vrea să treacă la acest echilibru alternativ.

Ce se va întâmpla cu structura ierarhică a cosmosului nostru dacă adăugăm la acest amestec tehnologia IA superinteligentă? Evident, tehnologia transportului şi comunicaţiilor se va îmbunătăţi spectaculos, aşa încât este de aşteptat ca această tendinţă istorică să persiste, cu noi niveluri ierarhice coordonând spaţii tot mai mari – cuprinzând, poate, în cele din urmă sisteme solare, galaxii, roiuri de galaxii şi mari regiuni din univers, după cum vom vedea în capitolul 6. Pe de altă parte, cea mai importantă forţă descentralizatoare va dăinui: este o risipă absurdă să coordonezi spaţii atât de vaste. Nici măcar Stalin n-a încercat să dicteze cu precizie când să meargă cetăţenii lui la baie. Pentru IA superinteligentă, legile fizicii vor impune limite superioare ferme asupra tehnologiei transportului şi comunicaţiilor, făcând improbabil ca nivelurile cele mai înalte ale ierarhiei să poată micromanageria tot ce se întâmplă la scară planetară sau locală. 0 IA superinteligentă din galaxia Andromeda nu-ţi va putea da ordine utile pentru deciziile tale de zi cu zi, fiindcă ar trebui să aştepţi instrucţiunile peste cinci milioane de ani (timpul dus-întors în care poţi schimba mesaje care călătoresc cu viteza

luminii). Timpul dus-întors pentru un mesaj care înconjoară Pământul este de aproximativ o,r secunde (ceea ce corespunde scării de timp la care au loc procesele noastre mentale), aşa încât un creier IA de dimensiunea Pământului poate avea gânduri globale cam în ritmul în care gândeşte creierul unui om. Pentru o mică inteligenţă artificială care efectuează o operaţiune la fiecare miliardime de secundă (ceea ce e tipic la calculatoarele de azi), 0,1 secunde ar fi resimţite ca patru luni pentru noi, deci micromanagementul la care ar fi supusă de o IA controlând planeta ar fi la fel de ineficient ca şi cum ai cere permisiunea pentru cele mai simple decizii prin scrisori transatlantice purtate de corăbii de pe vremea Iui Columb.

Această limită impusa de fizică asupra transferului de informaţii e o provocare pentru orice IA care vrea să ia în stăpânire lumea, ca să nu mai vorbim de univers. înainte să evadeze, Pro- meteu s-a gândit serios cum să evite fragmentarea minţii, aşa încât numeroasele lui module IA rulând pe calculatoare din lumea întreagă să se coordoneze şi să acţioneze ca o singură entitate. Aşa cum omegii s-au confruntat cu o problemă de control încercând să-l împiedice pe Prometeu să evadeze, Prometeu s-a confruntat cu o problemă de autocontrol când a încercat să se asigure că nici una dintre părţile lui nu se va revolta. Nu ştim deocamdată cât de mare poate fi un sistem controlat de o IA direct sau indirect, printr-un fel de ierarhie cooperantă – chiar dacă desprinderea rapidă îi oferă un avantaj strategic decisiv.

în concluzie, întrebarea cum va fi controlat un viitor superin- teligent e fascinant de complexă şi nu cunoaştem încă răspunsul. Unii susţin că se va tinde spre autoritarism, alţii că se va tinde spre o creştere a puterii individuale.

O idee des întâlnită în SF este că oamenii se vor contopi cu maşinile, fie prin ameliorarea tehnologică a corpurilor biologice pentru a deveni cyborgi (organisme cibernetice), fie prin uploadarea minţilor noastre pe maşini. în cartea sa *TheAge ofEm,* economistul Robin Hanson vorbeşte despre cum ar putea arăta o lume plină de uploa- dări (numite şi emulaţii – de aici em). Cred că uploadarea este capătul extrem al spectrului cyborgilor, în care tot ce rămâne din om e software-ul. Cyborgii născociţi de Hollywood pot fi vizibil mecanici, ca Borg din *Star Trek,* dar şi foarte asemănători cu oamenii, ca Terminator. Inteligenţa uploadărilor din filme poate fi la nivel uman, ca în episodul „White Christmas" din *Black Mirror,* dar şi mult dincolo de nivelul uman, ca în *Transcendente.*

Dacă superinteligenţă va apărea într-adevăr, tentaţia de a deveni cyborgi sau uploadări va fi foarte mare. în cartea sa din 1988 *Mind Children,* Hans Moravec spune: „Viaţa prelungită îşi pierde rostul dacă suntem condamnaţi s-o petrecem zgâindu-ne prosteşte la maşini ultrainteligente care încearcă să ne explice descoperirile lor tot mai spectaculoase în gânguritul de bebeluşi pe care-l putem noi pricepe," Deja mulţi oameni poartă ochelari, au proteze auditive, stimulatoare cardiace şi proteze ortopedice, iar substanţe medicinale circulă în sângele lor.

Continuarea naturală a acestei tendinţe, susţine Ray Kurzweil în cartea sa *The Singularity Is Near,* este folosirea nanoroboţilor, sistemelor cu biofeedback inteligent şi a altor tehnologii pentru a ne înlocui mai întâi sistemele digestiv şi endocrin, sângele şi inima pe la începutul anilor 2030, pentru ca în următoarele două decenii să ne upgradăm scheletul, pielea, creierul şi restul organismului. El anticipează că vom păstra probabil estetica şi importanţa afectivă a corpurilor umane, dar le vom reproiecta pentru a-şi modifica rapid înfăţişarea după dorinţă, atât fizic, cât şi în realitatea virtuală (graţie noilor interfeţe creier-computer). Moravec este de acord cu Kurzweil că cyborgizarea va

însemna mult mai mult decât simpla ameliorare a ADN-ului nostru: „Un supraom obţinut prin inginerie genetică ar fi doar un fel de robot de mâna a doua, proiectat cu handicapul că poate fi construit numai prin sinteza proteinelor controlată de ADN.“ El susţine apoi că va fi de preferat să eliminăm complet corpul uman şi să uploadăm minţi, creând o emulaţie a întregului creier în software. O asemenea uploadare poate trăi într-o realitate virtuală sau poate fi încorporată într-un robot capabil să meargă, să zboare, să înoate, să călătorească în spaţiu sau să facă orice e permis de legile fizicii, neîmpovărat de teama de moarte şi neîngrădit de resurse cognitive limitate.

Deşi asemenea idei par de domeniul SF, ele nu violează vreo lege cunoscută a fizicii, aşa încât întrebarea interesantă nu e dacă *pot* prinde viaţă, ci dacă nor prinde viaţă şi când. Unii cercetători cred că prima IAG la nivel uman va fi o uploadare, iar acesta va fi începutul drumului către superinteligenţă.*****

Trebuie totuşi să spun că majoritatea cercetătorilor IA şi a specialiştilor in neuroştiinţe nu împărtăşesc aceste idei, ci cred că cea mai rapidă cale spre superinteligenţă presupune ocolirea emulării creierului şi construirea ei într-un alt mod – după care s-ar putea să nu mai fim interesaţi de emularea creierului. în fond, de ce ar trebui ca drumul nostru cel mai simplu către o nouă tehnologie să fie acela urmat de evoluţie, constrâns de cerinţele de a se autoasambla, autorepara şi autoreproduce? Evoluţia optimizează eficienţa energetică din cauza resurselor limitate de hrană, dar nu optimizează simplitatea construcţiei sau înţelegerea de către inginerii umani. Meia, soţia mea, spune adesea că

***** După cum a explicat Bostrom, capacitatea de a simula un dezvoltator IA uman de vârf la un cost mult mai mic decât salariul Iui pe oră ar permite unei companii IA să-şi diminueze spectaculos forţa de muncă, îmbogăţin- du-se şi accelerând în mod recursiv progresul în construirea unor computere tot mai bune, iar în ultimă instanţă a unor minţi tot mai deştepte. (N. *a.)*

industria aviatică n-a început prin a fabrica păsări mecanice. în 2011, când în fine am reuşit să construim păsări mecanice,[1] la peste un secol după ce fraţii Wright au zburat pentru prima dată, industria aviatică nu s-a arătat interesată să treacă la transportul cu păsări mecanice care bat din aripi, deşi energetic e mai eficient – fiindcă vechea noastră soluţie e mai bine adaptată nevoilor noastre de a călători.

în acelaşi fel, bănuiesc că există căi mai simple de a construi maşini care să gândească la nivel uman decât soluţia găsită de evoluţie, iar dacă într-o bună zi vom reuşi să replicăm sau să uploadăm creiere, asta se va întâmpla abia după ce vom fi descoperit una dintre acele soluţii mai simple. Ea va consuma probabil mai mult decât cei 12 waţi folosiţi de creierul nostru, dar proiec- tanţii ei vor fi mai puţin obsedaţi de eficienţa energetică decât a fost evoluţia – iar curând vor putea utiliza maşinile lor inteligente pentru a proiecta maşini mai eficiente energetic.

CE SE VA ÎNTÂMPLA DE FAPT?

Evident, răspunsul scurt este că habar n-avem ce se va întâmpla dacă omenirea va reuşi să construiască o IAG la nivel uman. Acesta e motivul pentru care în capitolul de faţă am explorat un spectru larg de scenarii. Am încercat să cuprind întregul domeniu al speculaţiilor discutate de specialiştii în IA: desprindere rapidă/des- prindere lentă/absenţa desprinderii, control exercitat de oameni/ maşini/cyborgi, un centru de putere/mai multe centre de putere etc. Mi s-a spus despre cutare sau cutare lucru că nu se va întâmpla. Cred totuşi că în stadiul actual e înţelept să rămânem modeşti şi să recunoaştem cât de puţine ştim, căci fiecare scenariu prezentat mai sus este considerat o posibilitate reală de cel puţin unul dintre cercetătorii IA de vârf.

Cu trecerea timpului, pe măsură ce drumul urmat de noi va depăşi anumite bifurcaţii, vom începe să răspundem la între- bări-cheie şi să ne limităm opţiunile. Prima mare

întrebare este: „Vom reuşi vreodată să construim o IAG la nivel uman?“ Premisa acestui capitol este că vom reuşi, însă există experţi IA care cred că asta nu se va întâmpla, cel puţin nu în următoarele sute de ani. Timpul ne-o va spune! După cum am arătat mai sus, aproximativ o jumătate dintre experţii IA participanţi la conferinţa noastră din Puerto Rico au anticipat că aceasta se va întâmpla prin 2055. La o conferinţă organizată doi animai târziu, termenul a scăzut la 2047.

înainte de a construi o IAG la nivel uman am putea căpăta indicii care să ne spună dacă vom ajunge la această piatră de hotar prin inginerie informatică, uploadarea minţii sau vreo altă abordare deocamdată necunoscută. Dacă abordarea prin inginerie informatică, acum dominantă în rândul specialiştilor, va întârzia să ne ofere IAG vreme de secole, uploadarea va avea şanse mai mari să reuşească prima, aşa cum se întâmplă (cam nerealist) în filmul *Transcendence.*

Dacă ne vom apropia de IAG la nivel uman, vom avea indicii mai clare privind răspunsul la următoarea întrebare-cheie: „Va fi o desprindere rapidă, o desprindere lentă sau nu va fi nici o desprindere?“ După cum am văzut, o desprindere rapidă înlesneşte preluarea controlului asupra lumii, în vreme ce o desprindere lentă favorizează apariţia mai multor actori aflaţi în competiţie. Nick Bostrom disecă această problemă a vitezei de desprindere în- tr-0 analiză a ceea ce el numeşte *putere de optimizare şi recalcitranţă,* care în esenţă reprezintă cantitatea de efort preţios pentru a face ca IA să devină mai deşteaptă şi, respectiv, dificultatea de a obţine progrese. Rata medie a progresului creşte dacă e adusă în joc mai multă putere de optimizare, şi scade dacă ne confruntăm cu o recalcitranţă mai mare. El arată de ce recalcitranţa poate la fel de bine să crească sau să scadă când IAG ajunge Ia nivel uman şi îl depăşeşte. Puterea de optimizare însă e extrem de probabil să crească rapid când IAG depăşeşte nivelul uman, din motivele pe care le-am văzut în scenariul Omega: input-ul principal pentru

optimizarea ulterioară nu vine de la oameni, ci de la maşina însăşi, deci cu cât aceasta devine mai capabilă, cu atât se perfecţionează mai rapid (presupunând că recalcitranţa rămâne aproximativ constantă).

Pentru orice proces a cărui putere creşte cu o rată proporţională cu puterea curentă, rezultatul este că puterea lui se dublează continuu la intervale regulate. Numim *exponenţială* această creştere, iar asemenea procese le numim *explozii* Dacă numărul naşterilor creşte proporţional cu dimensiunea populaţiei, putem avea o explozie demografică. Dacă producerea de neutroni capabili să fisioneze atomii de plutoniu creşte proporţional cu numărul acestor neutroni, putem avea o explozie nucleară. Dacă inteligenţa maşinii creşte cu o rată proporţională cu puterea ei curentă, putem avea o explozie a inteligenţei. Toate aceste explozii sunt caracterizate de timpul în care îşi dublează puterea. Dacă pentru explozia inteligenţei timpul este de ordinul orelor sau zilelor, ca în scenariul Omega, avem de-a face cu o desprindere rapidă.

Ordinul de mărime al timpului caracteristic exploziei de pinde critic de ce anume e necesar pentru perfecţionarea IA: e suficient doar un nou software (care poate fi creat în secunde, minute sau ore) sau e nevoie de un nou hardware (care poate fi obţinut în luni sau ani). în scenariul Omega, exista un *exces de hardware,* conform expresiei lui Bostrom: omegii compensaseră slaba calitate a software-ului lor iniţial printr-o mare cantitate de hardware, ceea ce însemna că Prometeu îşi putea dubla de multe ori calitatea ameliorându-şi doar software-ul. Exista de asemenea un *exces de conţinut* sub forma bogăţiei datelor de pe intemet; Prometeu 1.0 nu era încă destul de deştept pentru a se folosi de cele mai multe dintre ele, dar odată ce inteligenţa lui a crescut, datele de care avea nevoie pentru a învăţa mai departe erau deja *disponibile.*

Costurile hardware-ului şi electricităţii sunt de asemenea importante, căci nu vom avea o explozie a inteligenţei până când costul unei munci la nivel uman nu scade sub salariul

pe oră al unui om. Să presupunem, de pildă, că prima IAG la nivel uman poate rula eficient pe norul Amazon la preţul de un milion de dolari pe oră de muncă la nivel uman. Ar fi fără îndoială o ştire de senzaţie, dar această IA n-ar putea trece prin autoperfecţionări recursive, fiindcă ar fi mult mai ieftin sa fie folosiţi în continuare oameni pentru a o perfecţiona. Să presupunem că aceşti oameni ar reuşi să scadă treptat costul la 100 ooo, io ooo, i ooo, 100, io dolari/oră, ajungând în cele din unnă la i dolar/oră. Când costul folosirii calculatorului pentru a se reprograma pe sine scade mult sub costul folosirii programatorilor umani pentru a-l «programa, oamenii pot fi lăsaţi deoparte, iar puterea de optimizare creşte enorm cumpărând timp de *cloud-computing.* Acest proces duce la scăderea în continuare a costurilor, sporind şi mai mult puterea de optimizare – şi astfel începe explozia inteligenţei.

Ajungem aşadar la ultima întrebare-cheie: „Cine sau ce va controla apoi explozia inteligenţei, şi care îi vor fi obiectivele?“ Vom examina posibilele obiective şi rezultate în capitolul următor şi, mai în detaliu, în capitolul 7. Pentru a lămuri problema controlului trebuie să ştim cât de bine poate fi controlată o IA şi cât de mult poate controla o IA.

în privinţa rezultatului final, opiniile sunt împărţite: unii cred că suntem condamnaţi, alţii susţin că el nu poate fi decât excepţional. Eu unul cred însă că în această dilemă se ascunde o capcană: este o greşeală să te întrebi pasiv „Ce se va întâmpla?“, de parcă ar exista vreo predestinare! Dacă nişte extratereştri superiori tehnologic ar sosi mâine, ar fi într adevăr cazul să ne întrebăm „Ce se va întâmpla?“ când navele lor s-ar apropia, fiindcă puterea lor ar fi probabil incomparabil mai mare decât a noastră, încât n-am putea influenţa cu nimic rezultatul. Pe de altă parte, dacă apare o civilizaţie tehnologic superioară bazată pe IA pentru că noi am construit-o, am putea influenţa rezultatul – influenţă exercitată atunci când creăm acea IA. Se cuvine deci să ne întrebăm: „Ce *ar trebui* să se întâmple? Ce viitor ne dorim?“ în capitolul 5 vom examina un spectru larg de consecinţe

posibile ale actualei curse către IAG, şi sunt curios să ştiu cum le clasificaţi, de la cea mai favorabilă la cea mai defavorabilă. Abia după ce ne vom fi gândit serios la ce fel de viitor ne dorim vom putea începe să ne croim drumul către un viitor dezirabil. Dacă nu ştim ce vrem, e puţin probabil să ajungem la el.

REZUMAT:

- Dacă într-o bună zi vom reuşi să construim o IAG la nivel uman, aceasta ar putea declanşa o explozie a inteligenţei care să ne lase mult în urmă.
- Dacă un grup de oameni reuşesc să controleze o explozie a inteligenţei, ei ar putea prealua controlul asupra lumii în câţiva ani.
- Dacă oamenii nu reuşesc să controleze o explozie a inteligenţei, IAînsăşi ar putea prelua controlul asupra lumii încă mai repede.
- în vreme ce o explozie rapidă a inteligenţei e probabil să conducă la o unică putere mondială, una lentă, durând ani sau decenii, e mai probabil să conducă la un scenariu multipolar, cu un echilibru de putere între mai multe entităţi oarecum independente.
- în istoria ei, viaţa s-a organizatîntr-o ierarhie tot mai complexă, determinată de colaborare, competiţie şi control. Superinteligenţa e probabil să permită coordonarea la scară cosmică, darnu-i limpede dacă va conduce în cele din urmă la un control totalitar de susîn jos sau la sporirea puterii individuale.
- Cyborgii şi uploadările sunt o posibilitate, dar pesemne că nu reprezintă calea cea mai rapidă către o inteligenţă avansată a maşinilor.
- Punctul culminant al actualei curse către IA ar putea fi atât cel mai bun, cât şi cel mai rău lucru pentru omenire, cu un spectru larg de rezultate posibile, despre care vom vorbi în următorul capitol.
- Trebuie să ne gândim serios la rezultatul pe care ni-l dorim şi la felul în care să ne croim drumul spre el, fiindcă dacă nu ştim ce vrem, e puţin probabil să ajungem la el.

CAPITOLUL 5

Consecinţe: următorii io ooo de ani

> *Este uşor şi ne închipuim gândurile omului eliberate de sclavia unui trup muritor – credinţa în viaţa de apoi e foarte răspândită. Dar nu-i nevoie să adoptăm o poziţie mistică sau religioasă pentru a accepta această posibilitate. Calculatoarele oferă un model chiar şi pentru cel mai înfocat mecanicist.* — Hans Moravec, *Mind Children*

> *Eu unul salut noua stăpânire a calculatoarelor.* — Ken Jennlngs, după ce a fostînfrânt la jocul *leopardy!* de calculatorul IBM Watson

> *Oamenii vor deveni la fel de neînsemnaţi ca gândacii.* — Marshall Brain

Cursa către IAG a început, şi nu ştim cum se va desfăşura, dar asta n-ar trebui să ne împiedice să reflectăm la consecinţele pe care ni le dorim, fiindcă ceea ce ne dorim influenţează rezultatul. Tu ce anume preferi, şi de ce?

1. Vrei să apară o superinteligenţă?
2. Vrei ca oamenii să existe în continuare, să fie înlocuiţi, cyborgizaţi şi/sau uploadaţi/simulaţi?
3. Controlul vrei să fie exercitat de oameni sau de maşini?
4. Vrei ca inteligenţele artificiale să fie conştiente sau nu?
5. Vrei să maximizezi experienţele pozitive, să minimizezi suferinţele sau să laşi ca asta să se aranjeze de la sine?
6. Vrei ca viaţa să se răspândească în cosmos?
7. Vrei o civilizaţie care să năzuiască spre un scop măreţ la care aderi, sau ai accepta ca formele de viaţă din viitor să se mulţumească cu obiective care ţi se par banale?

Scenarii ale consecinţelor IA:

Utopia libertarlană	**Oameni, cyborgi, uploadări şi superinteligenţe coexistă paşnic graţie drepturilor de proprietate.**
Dictator binevoitor	**Toată lumea ştie că IA conduce societatea şi impune reguli stricte, dar majoritatea oamenilor consideră că e un lucru bun.**
Utopia egalitară	**Oameni, cyborgi, uploadări şi superinteligenţe coexistă paşnic graţie abolirii proprietăţii şi venitului garantat.**
Gardian	**Este creată 0 IA superinteligentă cu scopul de a interveni discret pentru a împiedica crearea unei alte superinteligenţe. Ca urmare, abundă roboţii cu 0 inteligenţă uşor subumană, există şi cyborgi om-maşină,**
Zeu protector	**0 IA în esenţă omniscientă şi omnipotentă sporeşte la maximum fericirea oamenilor intervenind doar aşa încât să nu ne ştirbească senzaţia că suntem stăpâni pe destinul nostru, şi se ascunde atât de bine încât mulţi oameni chiar se îndoiesc de existenţa acestei IA.**
zeu înrobit	**0 IA superinteligentă e ţinută prizonieră de oameni care 0 folosesc pentru a produce tehnologii şi bogăţii de neînchipuit, în scopuri ce pot fi benefice sau malefice, după cum vor controlorii umani.**
învingători	**IA preia controlul, hotărăşte că oamenii sunt 0 ameninţare, 0 pacoste sau risipesc resursele, şi se debarasează de noi prin- tr-o metodă pe care nici măcar n-**
Descenaenp	**Inteligenţele artificiale îi înlocuiesc pe oameni, dar ne oferă 0 ieşire din scenă elegantă, făcându-nesă le consideram demni urmaşi ai noştri, cam la fel cum părinţii sunt bucuroşi şi mândri să aibă copii mai deştepţl decât ei, care învaţă de la ei, iar apoi înfăptuiesc ceea ce ei visau – chiar dacă nu pot trăi să vadă toate acestea.**
Paznic la zoo	**0 IA omnipotentă păstrează câţiva oameni care se simt trataţi ca animalele de la grădina zoologică şi îşi deplâng**
1984	**Progresul tehnologic către superinteligenţă e permanent controlat nu de 0 IA, ci de un stat orwellian al supravegherii condus de oameni, în care anumite tipuri de cercetări JA sunt interzise.**
Revenire	**Progresul tehnologic către superinteligenţă e împiedicat de revenirea la 0 societate pretehnologică, în genul comunităţii amish.†††††**
Autodistrugere	**Superinteligenţă nu ajunge să fie creată fiindcă omenirea se distruge prin alte mijloace (masacru nuclear şi/sau biotehno- logic provocat de schimbările climatice).**

Tabelul 5.1 Rezumat al scenariilor consecinţelor IA Pentru a alimenta asemenea consideraţii, vă propun să explorăm gama largă de scenarii rezumate în tabelul 5.1. Desigur, lista nu-i

††††† Grupare anabaptistă din Statele Unite, urmaşă a unor imigranţi elveţieni, care duce un trai decuplat de lumea modernă. *(N. t)*

exhaustivă, dar am ales-o pentru a acoperi întreg spectrul posibilităţilor. Nu vrem să ajungem într-o situaţie fără ieşire din cauza unei proaste planificări. Vă recomand să răspundeţi în scris la întrebările l-7 şi să reveniţi la răspunsuri după ce citiţi acest capitol, pentru a vedea dacă nu v-aţi răzgândit! 0 puteţi face la http://AgeOfAi.org, unde vă puteţi compara observaţiile şi discuta cu alţi cititori.

Scenariu	Există super-Inteligență?	Există oameni?	Oamenii dețin controlul?	Sunt oamenii în siguranță?	Sunt oamenii fericiți?	Există conștiință?
Utopia libertarians	Da	Da	Nu	Nu	Mixt	Da
Dictator binevoitor	Da	Da	Nu	Da	Mixt	Da
Utopia egalitar	Nu	Da	Da?	Da	Da?	Da
Gardian	Da	Da	Parțial	Posibil	Mixt	Da
Zeu protector	Da	Da	Parțial	Posibil	Mixt	Da
Zeu înrobit	Da	Da	Da	Posibil	Mixt	Da
învingători	Da	Nu	-	-	-	7
Descendenţl	Da	Nu	-	-	-	7
Paznic ia zoo	Da	Da	Nu	Da	Nu	□a
1984	Nu	Da	Da	Posibil	Mixt	Da
Revenire	Nu	Da	Da	Nu	Mixt	Da
Autodistrugere	Nu	Nu	-	-	-	Nu

Tabelul 5.2 Caracteristici ale scenariilor consecinţelor IA

UTOPIA LIBERTARIANĂ

Să începem cu un scenariu în care oamenii coexistă paşnic cu tehnologia, iar în unele cazuri se contopesc cu ea, după cum îşi închipuie mulţi futurologi şi autori de SF:

Viaţa pe Pământ (şi dincolo de el – vom reveni asupra acestei teme în capitolul următor) este mai diversă ca niciodată. Privind din satelit, poţi lesne distinge zonele

maşinilor, zonele mixte şi zonele ocupate doar de oameni. Zonele maşinilor sunt uzine enorme controlate de roboţi şi instalaţii de calcul lipsite de viaţă biologică, menite a folosi cât mai eficient fiecare atom. Deşi din afară par monotone şi mohorâte, zonele maşinilor sunt spectaculos de însufleţite în interior, cu experienţe uimitoare petrecându-se în lumi virtuale în timp ce calcule colosale descifrează tainele universului şi elaborează tehnologii inovatoare. Pământul găzduieşte multe minţi superinteligenţe care concurează între ele şi colaborează, toate adăpostite în zonele maşinilor.

Rezidenţii zonelor mixte sunt un amestec straniu de computere, roboţi, oameni şi hibrizi ale acestor trei categorii. După cum anticipează, de pildă, Hans Moravec şi Ray Kurzweil, mulţi oameni au corpuri upgradate tehnologic în cyborgi de diferite niveluri, iar unii au minţi uploadate în hardware-uri noi, estompând deosebirea dintre om şi maşină. Majoritatea fiinţelor inteligente n-au o formă fizică permanentă, ci există ca un software capabil să treacă instantaneu de la un computer la altul şi manifestându-se în lumea fizică prin corpuri robotice. Deoarece aceste minţi se pot cu uşurinţă duplica sau contopi, „dimensiunea populaţiei" variază neîncetat. Descătuşarea de substratul lor fizic face ca aceste fiinţe să capete o perspectivă diferită asupra vieţii: sunt mai puţin individualiste fiindcă pot lesne partaja între ele module de cunoaştere şi experienţă, iar subiectiv se simt nemuritoare fiindcă îşi pot crea cu uşurinţă copii de rezervă ale lor însele. într-un anumit sens, entităţile centrale ale vieţii nu sunt minţile, ci experienţele: experienţe uimitoare dăinuie pentru că sunt permanent copiate şi retrăite de alte minţi, în vreme ce experienţe banale sunt şterse de proprietarii lor pentru a elibera spaţiul de stocare.

Deşi majoritatea interacţiilor au loc în medii virtuale datorită comodităţii şi vitezei, multor minţi le place în continuare să activeze şi în corpuri fizice. De pildă, versiuni uploadate ale lui Hans Moravec, Ray Kurzweil şi Larry Page

şi-au făcut un obicei din a
crea pe rând realităţi virtuale pe care le explorează apoi împreună, dar din când în când se amuză să zboare împreună în lumea reală întrupaţi în roboţi asemănători păsărilor. Unii dintre aceşti roboţi care cutreieră străzile, cerul şi lacurile din zonele mixte sunt controlaţi în mod asemănător de oameni uploadaţi şi augmentaţi, care au ales să se întrupeze în zonele mixte fiindcă vor să trăiască şi printre oameni.

Dimpotrivă, în zonele locuite doar de oameni maşinile cu o inteligenţă generală la nivel uman sunt interzise, Ia fel şi organismele biologice perfecţionate tehnologic. Aici viaţa nu diferă mult de cea din prezent, cu excepţia faptului că e mai prosperă şi mai comodă: sărăcia a fost în mare parte eradicată şi există tratamente pentru aproape toate bolile de azi. Oamenii, aflaţi în minoritate, care au ales aceste zone trăiesc într-un plan al conştiinţei inferior faţă de cel al tuturor celorlalţi şi înţeleg puţin din ce fac minţile semenilor mai inteligenţi din alte zone, dar mulţi dintre ei sunt mulţumiţi cu viaţa pe care o duc.

Economia IA

Imensa majoritate a calculelor sunt efectuate în zonele maşinilor, aflate în principal în proprietatea numeroaselor IA superinteli- gente care locuiesc acolo. în virtutea inteligenţei şi tehnologiei lor superioare, nici o altă entitate nu le poate contesta puterea. Aceste IA au căzut de acord să coopereze şi să se coordoneze reciproc sub un sistem de guvernare libertarian care nu cunoaşte alte reguli decât protejarea proprietăţii private. Drepturile de proprietate se extind la toate entităţile inteligente, inclusiv oamenii, şi explică existenţa zonelor locuite numai de oameni. La început, grupuri de oameni s-au asociat şi au hotărât să interzică în zonele lor vânzarea proprietăţilor către non-umani.

Graţie tehnologiei lor, IA superinteligente au devenit mai bogate decât oamenii cu un factor mult mai mare decât cel dintre Bill Gates şi ultimul cerşetor. Cu toate acestea, oamenii din zonele exclusiv umane o duc mai bine decât majoritatea oamenilor de azi: economia lor e aproape decuplată de cea a maşinilor, aşa încât prezenţa maşinilor în alte locuri are o influenţă mică asupra lor, exceptând preluarea unor tehnologii pe care le pot înţelege şi re-produce – aşa cum comunitatea amish şi diferite triburi au un nivel de trai cel puţin la fel de ridicat ca în trecut. Nu contează că oamenii n-au ce să vândă maşinilor, din moment ce maşinile nu cer nimic în schimb.

în sectoarele mixte, diferenţa de bogăţie dintre inteligenţele artificiale şi oameni e vizibilă, ceea ce face ca pământul (singurul bun deţinut de oameni pe care maşinile vor să-l cumpere) ajunge la preţuri astronomice în comparaţie cu alte bunuri. Majoritatea oamenilor proprietari de pământ ajung aşadar să vândă o mică parte din el inteligenţelor artificiale, primind în schimb pe veci un venit de bază pentru ei şi urmaşii/uploadările lor. Aceasta i-a eliberat de nevoia de a munci şi i-a lăsat să se bucure de o uimitoare bogăţie de bunuri şi servicii ieftine produse de maşini, în realitatea fizică şi în cea virtuală. Din punctul de vedere al maşinilor, zonele mixte sunt mai curând pentru distracţie decât pentru muncă.

De ce ar fi imposibil acest scenariu

înainte de a ne lăsa seduşi de aventurile pe care le-am trăi ca cyborgi sau uploadări, să luăm în considerare unele motive pentru care acest scenariu ar fi imposibil. în primul rând, există două căi prin care am putea ajunge la forme umane perfecţionate (cyborgi sau uploadări):

1. Descoperim noi înşine cum să le creăm.
2. Construim maşini superinteligente care să facă asta în

locul nostru.

Dacă e urmată mai întâi calea i, ea conduce în mod firesc la o lume plină de cyborgi şi uploadări. După cum am văzut însă în capitolul precedent, majoritatea cercetătorilor IA cred că e mai probabil ca lucrurile să se petreacă invers, creierele îmbunătăţite sau digitale fiind mai greu de construit decât o IAG supraumană pornind de la zero – la fel cum păsările mecanice s-au dovedit a fi mai greu de construit decât avioanele. După apariţia unei maşini IA puternice, nu-i deloc evident că se vor mai crea vreodată cyborgi sau uploadări. Dacă neanderthalienii ar fi avut la dispoziţie încă 100 ooo de ani ca să evolueze şi să devină mai deştepţi, ei ar fi putut prospera – însă *Homo sapiens* nu le-a dat răgazul ăsta.

în al doilea rând, chiar dacă s-ar înfăptui scenariul cu cyborgi şi uploadări, nu-i deloc sigur că ar fi stabil şi ar dura. De ce ar rămâne stabil vreme de milenii echilibrul de putere dintre mai multe superinteligenţe, în loc ca inteligenţele artificiale să se contopească sau cea mai deşteaptă să preia controlul? Apoi, de ce ar respecta maşinile drepturile de proprietate ale oamenilor şi i-ar păstra pe aceştia în viaţă, odată ce n-au nevoie de oameni şi pot face ele însele mai bine şi mai ieftin toată munca oamenilor? Ray Kurzweil crede că oamenii naturali şi cei amelioraţi vor fi feriţi de exterminare fiindcă „oamenii vor fi respectaţi de inteligenţele artificiale pentru că ei au creat maşinile“.[1] După cum vom vedea însă în capitolul 7, nu trebuie să cădem în capcana antropomorfizării inteligenţelor artificiale şi să presupunem că simt emoţii de tip uman sau recunoştinţă. într-adevăr, deşi noi, oamenii, avem o înclinaţie spre gratitudine, nu manifestăm prea multă gratitudine faţă de creatorul intelectului nostru (ADN-ul) pentru a ne abţine să-i sabotăm obiectivele apelând la controlul natalităţii.

Chiar dacă am accepta că inteligenţele artificiale vor

alege să respecte drepturile de proprietate ale oamenilor, ele pot obţine treptat mult din teritoriul nostru pe alte căi, folosind puterile lor superinteligenţe de persuasiune, despre care am vorbit în capitolul precedent, pentru a-i convinge pe oameni să vândă pământ în schimbul unei vieţi îmbelşugate. în sectoarele exclusiv umane ele i-ar putea ispiti pe oameni să lanseze campanii politice pentru a permite vânzarea de terenuri. La urma urmei, până şi bio-ludiţi convinşi ar fi dispuşi să vândă pământ pentru a salva viaţa unui copil bolnav sau pentru a dobândi nemurirea. Dacă oamenii sunt educaţi, au parte de distracţii şi sunt ocupaţi, declinul natalităţii ar putea duce la scăderea populaţiei lor fără intervenţia maşinilor,

aşa cum se întâmplă acum în Japonia şi Germania. Oamenii ar putea dispărea astfel în câteva mii de ani.

Neajunsuri

Pentru unii dintre cei mai înflăcăraţi partizani ai lor, cyborgii şi uploadările oferă speranţa fericirii tehnologice şi prelungirii vieţii pentru toată lumea. într-adevăr, perspectiva de a fi uploadat în viitor a îndemnat peste o sută de oameni să aleagă îngheţarea postumă a creierului lor de către compania Alcor din Arizona. Dacă însă această tehnologie va fi disponibilă în viitor, nu-i deloc sigur că va fi la îndemâna oricui. Probabil că mulţi dintre cei foarte bogaţi se vor bucura de ea, dar ceilalţi? Chiar dacă tehnologia s-ar ieftini, unde s-ar trage linia de demarcaţie? Ar fi uploadate creiere grav deteriorate? Am uploada fiecare gorilă? Fiecare furnică? Fiecare plantă? Fiecare bacterie? Ar acţiona civilizaţia viitorului ca un colecţionar obsesiv-compulsiv încercând să uploadeze totul, sau măcar câteva exemplare interesante din fiecare specie, ca pentru o arcă a lui Noe? Poate doar câteva exemplare reprezentative din fiecare tip uman? Pentru entităţile incomparabil mai inteligente care ar exista la acea vreme, un om uploadat ar părea la fel de interesant cum ne-ar părea nouă un şoarece sau un melc simulat. Deşi din punct de vedere tehnic putem readuce la viaţă vechile programe informatice din anii 1980 pe un emulator DOS, cei mai mulţi dintre noi nu găsesc că e suficient de interesant pentru a o face efectiv.

Unora le-ar displăcea acest scenariu utopic libertarian pentru că nu înlătură suferinţa care poate fi evitată. Odată ce singurul principiu sacru este dreptul de proprietate, nimic nu împiedică perpetuarea suferinţelor din lumea de azi în zonele umane şi mixte. în timp ce unii oameni ar prospera, alţii ar putea ajunge să trăiască în mizerie şi servitute, ori să sufere din pricina violenţei, fricii, represiunii sau depresiei. De pildă, romanul din 2003 *Manna*

al lui Marshall Brain arată cum progresul IA într-un sistem economic libertarian face ca majoritatea americanilor să devină neangajabili şi îi condamnă să-şi trăiască restul vieţii în aziluri triste administrate de roboţi. La fel ca animalele din ferme, ei sunt hrăniţi şi îngrijiţi în spaţii înguste, unde cei bogaţi nu vin niciodată să-i vadă. Substanţe sterilizante dizolvate în apă îi împiedică să se înmulţească, aşa încât populaţia scade treptat, iar cei bogaţi se bucură în măsură tot mai mare de avuţia produsă de roboţi.

în scenariul utopiei libertariene, suferinţa nu se limitează neapărat la oameni. Dacă unele maşini au experienţe emoţionale conştiente, pot şi ele suferi. De exemplu, un psihopat răzbunător ar putea avea dreptul să ia o copie uploadată a duşmanului său şi să-i aplice cele mai cumplite torturi într-o lume virtuală, provo- cându-i dureri de o intensitate şi o durată mult dincolo de ceea ce este biologic posibil în lumea reală.

DICTATOR BINEVOITOR

Să examinăm acum scenariul în care toate aceste forme de suferinţă sunt absente pentru că o unică superinteligenţă binevoitoare conduce lumea şi impune reguli stricte menite a maximiza ceea ce ea consideră a fi fericirea umană. Acesta e unul dintre rezultatele posibile ale primului scenariu Omega din capitolul precedent, în care omegii cedează controlul lui Prometeu după ce au înţeles cum îl pot face să-şi dorească o societate umană înfloritoare.

Mulţumită tehnologiilor excepţionale elaborate de IA dictatoare, omenirea a scăpat de sărăcie, boli şi alte probleme ţinând de înapoierea tehnologică, iar toţi oamenii se bucură de o viaţă prosperă şi tihnită. Toate nevoile lor de bază sunt satisfăcute, în vreme ce maşinile controlate de IA produc toate bunurile şi serviciile necesare. Infracţiunile sunt practic eliminate, fiindcă LA dictatoare e în esenţă

omniscientă şi pedepseşte pe oricine încalcă regulile. Toţi poartă brăţara electronică pomenită în capitolul precedent (sau o mai comodă versiune a ei implantată) care îi supraveghează în timp real, îi pedepseşte, sedează şi execută. Toţi ştiu că trăiesc într-o dictatură IA poliţienească, dar cei mai mulţi consideră că e un lucru bun.

IA dictatoare superinteligenţă are ca obiectiv să descopere cum arată utopia umană, date fiind preferinţele codificate în genele noastre, şi s-o pună în aplicare. Graţie prevederii înţelepte a celor care au creat-o, IA nu încearcă pur şi simplu să ne maximizeze fericirea conform propriilor noastre declaraţii – punând, de pildă, tuturor perfuzii cu morfină –, ci foloseşte o definiţie subtilă şi complexă a prosperităţii umane, transformând Pământul într-un mediu bogat biologic în care e o bucurie să trăieşti. Ca urmare, majoritatea oamenilor găsesc că viaţa lor e împlinită şi are un sens.

Sistemul sectoarelor

Respectând diversitatea şi recunoscând că oameni diferiţi au preferinţe diferite, IA a împărţit Pământul în sectoare pentru ca oamenii să aibă de unde alege, să se bucure de compania spiritelor înrudite. Iată câteva exemple:

- Sectorul cunoaşterii: Aici IA oferă o educaţie optimizată, inclusiv experienţe captivante în realitatea virtuală, permiţându-ţi să înveţi tot ce poţi în domenii la alegerea ta. Dacă vrei, anumite descoperiri spectaculoase nu ţi se vor prezenta, dar vei fi călăuzit să Ie redescoperi tu însuţi.
- Sectorul artelor: Ai aici ocazia să savurezi, să creezi şi să împărtăşeşti muzică, arte plastice, literatură şi alte forme de expresie artistică.
- Sectorul hedonist Locuitorii Iui îl numesc sectorul petrecerilor şi e tot ce-şi pot dori cei care tânjesc după

mâncăruri rafinate, pasiune, intimitate sau pur şi simplu distracţie fără limite.

■ Sectorul evlavios: Există de fapt mai multe asemenea sectoare, corespunzând diferitelor religii, ale căror reguli sunt impuse cu stricteţe.

• Sectorul naturii sălbatice: Dacă eşti în căutare de plaje superbe, lacuri romantice, munţi semeţi sau fiorduri fantastice, aici e locul tău.

• Sectorul jocurilor: Dacă îţi plac jocurile pe calculator, IA are pentru tine o ofertă fabuloasă.

• Sectorul virtual; Dacă vrei să te eliberezi de corpul tău fizic şi să pleci în vacanţă, IA îl va păstra hidratat, hrănit, exersat şi curat, în vreme ce tu vei explora lumi virtuale prin implanturi neuronale.

• Sectorul închisorilor: în cazul în care încâlci regulile, vei ajunge aici pentru a fi reeducat, dacă nu cumva vei primi instantaneu pedeapsa cu moartea.

în plus faţă de aceste sectoare tematice „tradiţionale“, există şi altele cu teme modeme, pe care oamenii de azi nici măcar nu le-ar putea înţelege. La început, oamenii sunt liberi să se mute dintr-un sector în altul când vor, ceea ce se face rapid graţie sistemului de transport hipersonic al IA. De pildă, după ce ai petrecut o săptămână intensă în sectorul cunoaşterii, aflând legile ultime ale fizicii descoperite de IA, ai putea hotărî să-ţi dai frâu liber în weekend în sectorul hedonist, iar apoi să te relaxezi câteva zile pe o plajă din sectorul naturii sălbatice.

IA impune două seturi de reguli: universale şi locale. Regulile universale se aplică în toate sectoarele – de pildă, interdicţia de a face rău altor oameni, de a producere armament sau de a încerca să creezi o superinteligenţă rivală. Sectoarele individuale mai au şi legi locale, codificând anumite valori morale. Sistemul sectoarelor permite aşadar tratarea diferenţiată a valorilor care nu sunt compatibile. Cele mai multe legi locale se aplică în sectorul

închisorilor şi în unele sectoare religioase, însă există un sector libertarian ai cărui rezidenţi se mândresc că n-au nici o regulă locală. Toate pedepsele, chiar şi cele locale, sunt puse în aplicare de IA, din moment ce pedepsirea unui om de către alt om ar încălca regula universală de a nu face rău. Dacă încâlci o regulă locală, IA te lasă să alegi (în cazul în care nu te afli în sectorul închisorilor) între a suporta pedeapsa prescrisă sau a fi exilat pe veci din acel sector. De pildă, dacă între două femei se naşte o relaţie amoroasă într-un sector unde homosexualitatea e pedepsită cu închisoarea (ca în multe ţări de azi), IA le lasă să aleagă între a merge la puşcărie sau a părăsi pentru totdeauna acel sector, iară să-şi mai întâlnească vreodată vechii prieteni (dacă nu cumva şi aceştia părăsesc sectorul).

Indiferent de sectorul unde se nasc, toţi copiii primesc o educaţie minimă de bază din partea IA, care cuprinde cunoştinţe despre omenire în ansamblul ei, spunându-li-se că sunt liberi să viziteze alte sectoare şi să se mute unde vor.

IA a proiectat un mare număr de sectoare în parte fiindcă a fost creată pentru a respecta diversitatea umană existentă azi. Fiecare sector oferă însă mai multă bunăstare decât ar permite tehnologia actuală, fiindcă IA a eliminat toate problemele tradiţionale, inclusiv sărăcia şi infracţiunile. De exemplu, oamenii din sectorul hedonist nu sunt îngrijoraţi de bolile cu transmitere sexuală (au fost eradicate), mahmureală sau dependenţă (IA a creat droguri perfecte, fără efecte secundare negative). Nici un om din nici un sector nu-şi pune problema bolilor, fiindcă IA poate repara corpurile umane cu ajutorul nanotehnologiei. Locuitorii a nume roase sectoare se bucură de arhitectura de înaltă tehnologie, faţă de care imaginaţia SF păleşte.

în rezumat, dacă atât scenariul utopiei libertariene, cât şi cel al dictatorului binevoitor implică tehnologie produsă de IA şi avuţie la niveluri excepţionale, ele se deosebesc în privinţa obiectivelor şi a celui care deţine comanda. în utopia libertariană, proprietarii tehnologiei hotărăsc ce să

facă cu ea, în vreme ce în acest scenariu IA dictatoare are puteri nelimitate şi stabileşte obiectivul suprem: transformarea Pământului într-o croazieră de plăcere all-inclusive tematizată în funcţie de preferinţele oamenilor. Din moment ce IA îi lasă pe oameni să aleagă între numeroase căi pentru atingerea fericirii şi se îngrijeşte de nevoile lor materiale, dacă cineva suferă, asta se întâmplă doar pentru că aşa a ales.

Neajunsuri

Deşi dictatura binevoitoare aduce cu sine o sumedenie de expe rienţe pozitive şi elimină practic suferinţa, mulţi oameni simt că lucrurile ar putea sta mai bine. în primul rând, unii îşi doresc ca oamenii să aibă mai multă libertate în făurirea societăţii şi a destinului lor, dar păstrează aceste dorinţe în sinea lor, fiindcă ştiu că ar fi o sinucidere să înfrunţi puterea copleşitoare a maşinii care îi conduce pe toţi. Unii nu vor să li se impună numărul copiilor pe care îi fac, şi îi deranjează faptul că IA insistă asupra obţinerii sustenabilităţii prin controlul naşterilor. Amatorii de arme detestă interdicţia de a fabrica şi folosi arme, iar unii oameni de ştiinţă sunt nemulţumiţi de interdicţia de a construi propria lor superinteligenţă. Mulţi sunt indignaţi de ce se întâmplă în alte sectoare, se tem că copiii lor vor alege să se mute acolo şî ar vrea să-şi impună pretutindeni codul lor moral.

Cu timpul, tot mai mulţi oameni aleg să se mute în acele sectoare unde IA le oferă practic orice experienţe îşi doresc. Spre deosebire de paradisul tradiţional, în care primeşti ce meriţi, acesta seamănă cu „Noul Paradis“ din romanul lui Julian Bames *O istorie a lumii în io şi Vi capitole* (şi cu Nice Place to Visit“ din serialul *Zona crepusculara),* unde primeşti ce-ţi doreşti. Paradoxal, destui oameni sfârşesc prin a se plânge că primesc ce vor. În povestea lui Bames, protagonistul se lasă un timp îndelungat în voia dorinţelor sale, de la mese copioase şi golf la sex cu celebrităţi, dar în

cele din urmă cade pradă plictisului şi cere anihilarea. în dictatura binevoitoare, mulţi au aceeaşi soartă, vieţile lor sunt plăcute, însă în cele din urmă lipsite de sens. Deşi îţi poţi crea provocări artificiale, de la redescoperirea ştiinţei Ia alpinism, toţi ştiu că provocarea nu-i reală, e doar o formă de amuzament. Este inutil ca oamenii să încerce să facă ştiinţă sau să descopere alte lucruri, din moment ce IA a făcut-o deja. Este inutil ca oamenii să încerce să creeze ceva care le îmbunătăţeşte viaţa, din moment ce e de-ajuns să ceară asta inteligenţei artificiale pentru a o obţine imediat.

UTOPIA EGALITARĂ

Ca un contrapunct la această dictatură lipsită de provocări, să examinăm acum un scenariu în care nu există o IA superinteligenţă, iar oamenii sunt stăpânii destinului lor. Aceasta e „civilizaţia de generaţia a patra“ din romanul lui Marshall Brain *Marna.* Este

antiteza economică a utopiei libertariene, în sensul că oamenii, cyborgii şi uploadările coexistă paşnic graţie nu drepturilor de proprietate, ci abolirii proprietăţii şi garantării veniturilor.

Viaţa în absenţa proprietăţii

Ideea centrală e împrumutată de la mişcarea *open-source software* (software cu sursă deschisă): dacă un software poate fi copiat gratis, atunci oricine îl poate folosi cât are nevoie, iar chestiunea posesiei şi a proprietăţii devine irelevantă‡‡‡‡ Conform legii cererii şi ofertei, costul reflectă raritatea unui bun, deci dacă oferta e practic nelimitată, preţul devine neglijabil. în acest spirit, toate drepturile de proprietate intelectuală sunt abolite: nu există brevete, copyright sau proiecte marcă înregistrată – oamenii îşi împart pur şi simplu între ei ideile bune, iar fiecare e liber să le folosească.

Graţie roboticii avansate, aceeaşi idee de absenţă a proprietăţii se aplică nu doar produselor bazate pe informaţie, cum sunt cărţile, filmele sau proiectele, ci şi produselor materiale, cum sunt casele, automobilele, hainele şi calculatoarele. Toate aceste produse nu sunt decât atomi aranjaţi în anumite moduri, şi nu ducem lipsă de atomi, aşa încât atunci când cineva vrea un produs, o reţea de roboţi va folosi unul dintre proiectele cu sursă deschisă disponibile ca să-l construiască gratuit pentru el. Se încurajează utilizarea materialelor uşor reciclabile, iar dacă cineva nu mai are nevoie de un obiect, roboţii pot rearanja atomii obiectului în ceva de care altcineva are nevoie. Toate resursele sunt astfel reciclate, nimic nu e permanent distrus. De asemenea, aceşti roboţi construiesc şi întreţin

‡‡‡‡‡ Ideea apare şi la Sfântul Augustin: „Căci orice lucru care nu îţi scade dându-l nu îl ai încă – atâta timp cât îl ai şi nu îl dai – aşa cum trebuie avut“ |De *doctina christiana,* Editura Humanitas, Bucureşti, 2002 – *n.* t]. *(N. a.)*

suficiente centrale pe baza energiei regenerabile (solară, eoliană etc.), prin urmare şi energia e practic gratuită.

Pentru a evita ca unii oameni lacomi să ceară atât de multe produse şi atât de mult pământ, încât alţii să rămână nevoiaşi, fiecare persoană primeşte de la guvern un venit lunar pe care-I poate cheltui după voie pentru produse şi pentru a-şi închiria o locuinţă. Nu ai motive să încerci să câştigi mai mulţi bani, odată ce venitul garantat e suficient ca să ţi acoperi nevoile rezonabile. Ar fi şi inutil, fiindcă ai de-a face cu oameni care îşi oferă gratuit produsele intelectuale şi cu roboţi care produc bunuri materiale la preţuri derizorii

Creativitate şi tehnologie

Drepturile de proprietate intelectuală sunt adesea considerate sursa creativităţii şi invenţiilor. Marshall Brain arată însă că multe dintre cele mai strălucite exemple de creativitate umană – de la descoperiri ştiinţifice la creaţie literară, artistică şi muzicală – n-au fost stimulate de dorinţa de profit, ci de alte porniri umane, cum sunt curiozitatea, imboldul de a crea sau nevoia de primi recunoaşterea semenilor. Nu banii i-au făcut pe Einstein să descopere relativitatea şi pe Linus Torvalds să creeze sistemul de operare gratuit Linux. Dimpotrivă, astăzi mulţi oameni nu reuşesc să-şi împlinească potenţialul creator fiindcă trebuie să-şi dedice timpul şi energia unor activităţi mai puţin creatoare, doar pentru a-şi câştiga existenţa. Eliberându-i pe savanţi, artişti, inventatori şi proiectanţi de corvezile lor şi permiţându-le să creeze din dorinţă autentică, societatea utopică a lui Marshall Brain se bucură de niveluri ale inovaţiei mai înalte decât cele de azi – şi, în consecinţă, de o tehnologie şi un standard de viaţă superioare.

O asemenea nouă tehnologie este o formă de hiperintemet numită Vertebrane. Ea îi conectează wireless pe toţi doritorii prin implanturi neuronale, oferindu-le acces

mental instantaneu la informaţiile gratuite din lumea întreagă doar pe calea gândirii, îţi permite să uploadezi orice experienţe vrei pentru a le împărtăşi, aşa încât ele pot fi retrăite de alţii, şi să înlocuieşti experienţele oferite de simţuri cu experienţe virtuale downloadate la alegere. *Mama* explorează numeroasele avantaje decurgând de- aici, inclusiv acela de a face cu uşurinţă exerciţii fizice:

> Cea mai mare problemă cu exerciţiile fizice epuizante este că nu-s deloc plăcute. Sunt dureroase. [...] Atleţii suportă durerea, dar majoritatea oamenilor obişnuiţi nu vor să îndure suferinţa vreme de-o oră sau mai mult. Aşa încât [...] cineva a găsit o soluţie. Iţi deconectezi creierul de input-ul senzorial şi priveşti un film, stai de vorbă, scrii mesaje, citeşti o carte sau faci orice ai chef vreme de-o oră. în acest timp, sistemul Verte- brane face exerciţii cu trupul tău, care trece printr-un ciclu aerobic complet, mai obositor decât ar suporta majoritatea oamenilor. Tu nu simţi nimic, dar corpul tău rămâne în formă.

în plus, calculatoarele din sistemul Vertebrane pot monitoriza input-ul senzorial al oricui şi pot dezactiva controlul lui motor dacă este pe punctul de a comite o infracţiune.

Neajunsuri

O obiecţie la această utopie egalitară este că defavorizează inteligenţa non-umană: roboţii care efectuează practic toată munca par să fie destul de inteligenţi, dar sunt trataţi ca nişte sclavi, iar oamenii consideră că aceştia sunt lipsiţi de conştiinţă şi n-ar trebui să aibă vreun drept Dimpotrivă, utopia libertariană garantează drepturi tuturor entităţilor inteligente, fără a favoriza tipul nostru bazat pe carbon. Odinioară, populaţia albă din sudul Statelor Unite a prosperat pe seama muncii sclavilor, dar astăzi pentru cei

mai mulţi dintre noi ar fi inacceptabil moral să numim asta progres.

Alt punct slab al scenariului utopiei egalitare este că ar putea fi instabil şi n-ar rezista pe termen lung, conducând spre unul dintre celelalte scenarii atunci când neabătutul progres tehnologic creează în cele din urmă o superinteligenţă. Din motive neexplicate în *Mama,* superinteligenţă încă nu există, iar noile tehnologii sunt inventate tot de oameni, nu de calculatoare. Cartea scoate totuşi în evidenţă tendinţe în această direcţie. De pildă, continuu perfecţionat, Vertebrane ar putea deveni superinteligent. Apoi, există un grup numeros de oameni numiţi Vites care aleg să-şi ducă viaţa aproape în întregime în lumea virtuală. Vertebrane are grijă de toate aspectele fizice în ce-i priveşte, inclusiv hrănirea, spălarea şi folosirea toaletei, activităţi de care mintea lor, trăind în realitatea virtuală, nu e conştientă. Pe aceşti Vites nu-i interesează să aibă copii fizici, iar ei mor pe rând cu corpurile lor fizice, aşadar, dacă toţi devin Vites, omenirea dispare într-o stare de beatitudine.

Vites consideră că trupul uman le distrage atenţia, iar noile tehnologii în curs de elaborare promit să elimine această pacoste, permiţându-le să trăiască mai mult sub formă de creiere fără trup, alimentate cu nutrienţi optimi. De aici următorul pas firesc şi dezirabil pentru Vites e să scape complet de creiere prin uploa- dare, prelungindu-şi astfel durata vieţii. Acum însă au dispărut toate limitările impuse de creier asupra inteligenţei, şi nu e clar ce ar sta în calea sporirii capacităţii cognitive a Vites până să se ajungă la autoameliorare recursivă şi la explozia inteligenţei.

GARDIAN

Am văzut că trăsătura atractivă a scenariului utopiei

egalitare este că oamenii sunt stăpânii propriului lor destin, dar şi că scenariul ar putea conduce spre distrugerea acestei trăsături prin dezvoltarea superinteligenţei. Lucrurile ar putea fi îndreptate prin construirea unui *gardian,* o superinteligenţă având obiectivul de a interveni cât de puţin e necesar pentru a împiedica crearea unei alte superinteligenţe.[§§§§§] Asta ar permite oamenilor să controleze mai departe utopia lor egalitară, pesemne chiar şi atunci când viaţa se răspândeşte în cosmos, după cum vom vedea în capitolul următor.

Cum ar fi cu putinţă? Gardianul IA ar avea încorporat acest obiectiv simplu, aşa încât să-l păstreze în timp ce suferă autoper- fecţionări recursive şi devine superinteligent. Ar folosi atunci tehnologia de supraveghere cea mai discretă şi mai neinvazivă pentru a monitoriza orice tentativă umană de a crea o superinteligenţă rivală. Ar împiedica apoi asemenea tentative în modul cel mai puţin perturbator. Pentru cei aflaţi abia la început ar iniţia şi răspândi meme[******] culturale lăudând virtuţile autodeterminării umane şi promovând evitarea superinteligenţei. Dacă totuşi unii cercetători ar căuta să obţină superinteligenţa, ar putea încerca să-i descurajeze. Dacă n ar izbuti, i-ar putea pune pe piste false şi, eventual, le-ar sabota eforturile. Cu accesul lui practic nelimitat la tehnologie, sabotajul gardianului ar trece neobservat – de pildă, dacă ar folosi nanotehnologia pentru a şterge discret din creierele cercetătorilor (şi din computere) amintirile privind progresele lor.

Hotărârea de a construi un gardian IA ar fi probabil controversată. Printre adversarii ei s-ar afla mulţi oameni religioşi care ar obiecta la ideea de a crea o IA

§§§§§ Ideea mi-a fost sugerată de prietenul şi colegul meu Anthony Aguirre. (Mu.)

****** Memă – termen creat de Richard Dawkins prin analogie cu gena; unitate de informaţie care se propagă de la o minte la alta. *(N.* t.)

superinteligentă cu puteri de ordin divin, afirmând că există deja un Dumnezeu şi ar fi o blasfemie să încerci să creezi o divinitate presupus superioară. Susţinătorii ideii ar putea afirma că gardianul nu doar ar menţine omenirea stăpână pe destinul ei, dar ar şi proteja-o de alte riscuri la care ar expune-o superinteligenţa, ca în scenariile apocaliptice discutate în capitolul precedent.

Pe de altă parte, criticii ar putea considera că existenţa unui gardian e un lucru îngrozitor, paralizând irevocabil potenţialul omenirii şi blocând pe veci progresul tehnologic. De pildă, dacă s-ar dovedi că răspândirea vieţii în cosmos necesită ajutorul unei superinteligenţe, gardianul ne-ar anula această imensă şansă lăsân- du-ne pentru totdeauna captivi în sistemul nostru solar. în plus, spre deosebire de zeii celor mai multe religii, gardianul IA este cu desăvârşire indiferent faţă de acţiunile oamenilor, atâta vreme cât ei nu creează o altă superinteligenţă. De exemplu, nu ne-ar împiedica să ne provocăm mari suferinţe sau chiar să ne autodistrugem, dispărând complet

ZEU PROTECTOR

Dacă vrem să folosim un gardian IA pentru a-i menţine pe oameni stăpâni pe propriul destin, am putea pesemne duce lucrurile mai departe, făcând ca această IA să aibă grijă de noi în mod discret, acţionând ca un zeu protector. în acest scenariu, IA superinteli- gentă este practic omniscientă şi omnipotentă, maximizând fericirea umană doar prin intervenţii care ne dau mai departe senzaţia că noi ne controlăm soarta şi ascunzându-se atât de bine, încât mulţi oameni se îndoiesc de existenţa ei. Cu excepţia camuflării, seamănă cu scenariul „Nanny AI“ („bona IA“) propus de cercetătorul Ben Goertzel.[2]

Zeul protector şi dictatorul binevoitor sunt deopotrivă „IA prietenoase“ care încearcă să sporească fericirea oamenilor, dar acordă prioritate unor nevoi umane diferite. Psihologul american Abraham Maslow a stabilit o celebră

ierarhie a nevoilor umane. Dictatorul binevoitor se îngrijeşte de nevoile aflate la baza ierarhiei, cum sunt hrana, adăpostul, siguranţa şi diverse plăceri. Zeul protector însă nu încearcă să maximizeze fericirea umană în sensul îngust de satisfacere a nevoilor noastre de bază, ci într-un sens mai profund, făcându-ne să simţim că viaţa noastră are un rost şi un scop. El urmăreşte să ne satisfacă toate nevoile, cu condiţia să rămână ascuns, lăsându-ne pe noi (în majoritatea cazurilor) să decidem.

Un zeu protector ar fi rezultatul firesc al primului scenariu Omega din capitolul precedent, în care omegii cedează controlul lui Prometeu, care în cele din urmă se ascunde şi elimină toate cunoştinţele oamenilor privind existenţa lui. Cu cât progresează tehnologia IA, cu atât e mai uşor pentru el să se ascundă. Filmul *Transcendente* dă un asemenea exemplu, în care nanomaşinile sunt omniprezente şi devin o parte naturală a lumii.

Monitorizând atent toate activităţile umane, zeul protector IA poate interveni cu mici impulsuri sau miracole ici-colo, împin- gându-ne pe o cale mult mai bună. De pildă, dacă ar fi existat în anii 19 30, ar fi putut pesemne aranja ca Hitler să moară de un atac cerebral, odată ce a înţeles care îi sunt intenţiile. Dacă ar părea că ne îndreptăm spre un război nuclear accidental, l-ar evita printr-o intervenţie pe care am considera-o noroc pur. Ne-ar putea de asemenea oferi „revelaţii“ sub forma ideilor pentru noi tehnologii benefice, inoculate în taină în vreme ce dormim.

Acest scenariu ar putea fi pe placul multor oameni, fiindcă seamănă cu credinţele şi speranţele din religiile monoteiste de azi. Dacă cineva îi pune superinteligenţei întrebarea „Există Dumnezeu?“, aceasta ar putea răspunde în glumă, ca Stephen Hawking: „Acum există!” Pe de altă parte, unii oameni religioşi ar putea dezaproba acest scenariu pentru că IA încearcă să-I întreacă în dumnezeire

pe zeul lor sau să se amestece într-un plan divin în care oamenii ar trebui să facă binele doar prin alegere personală.

Alt neajuns al acestui scenariu este că zeul protector lasă să apară unele suferinţe ce pot fi împiedicate, pentru ca existenţa lui să nu devină manifestă. Este analogul situaţiei din filmul *The Imitatwn Game,* în care Alan Turing şi colegii săi spărgători de coduri de la Blatchley Park†††††† aflau dinainte de atacurile submarinelor germane asupra convoaielor navale ale Aliaţilor, dar alegeau să intervină doar într-un număr mic de cazuri, pentru a nu deconspira avantajul secret de care dispuneau. Este interesant de comparat această situaţie cu aşa-numita problemă a *teodiceei:* de ce îngăduie un zeu bun existenţa suferinţei? Unii teologi au susţinut că Dumnezeu vrea să le lase oamenilor o anumită libertate. în scenariul zeului protector IA, răspunsul la problema teodiceei este că libertatea percepută îi face în ansamblu pe oameni mai fericiţi

Un al treilea neajuns al acestui scenariu este că oamenii se bucură de un nivel tehnologic mult mai scăzut decât cel la care a ajuns IA superinteligentă. Dacă un dictator binevoitor poate întrebuinţa întreaga tehnologie pe care a inventat-o în beneficiul omenirii, un zeu protector e limitat de capacitatea oamenilor de a reinventa (primind indicii subtile) şi înţelege tehnologia lui. El ar putea de asemenea limita progresul tehnologic uman pentru a se asigura că propria lui tehnologie rămâne cu mult superioară, ca să nu fie detectată.

---

†††††† în timpul celui de-al Doilea Război Mondial, în conacul de la Bletchley Park (în Buckinghamshire, nu departe de Londra), britanicii au adus o echipă de spărgători de coduri care, sub conducerea lui Alan Turing şi folosind □ formă primitivă de calculator, a reuşit să descifreze comunicabile secrete ale armatei germane, ceea ce a jucat un rol important în câştigarea războiului. (IV. 1)

N-ar fi oare minunat dacă am putea combina caracteristicile cele mai atrăgătoare ale tuturor scenariilor de mai sus, folosind tehnologia elaborată de superinteligenţă pentru a elimina suferinţa, dar rămânând în acelaşi timp stăpâni pe destinul nostru? Aşa arată scenariul zeului înrobit, în care o IA superinteligentă se află sub controlul oamenilor pentru a produce tehnologie şi bogăţii inimaginabile. Scenariul Omega de la începutul cărţii se încheie în acest fel dacă Prometeu nu e eliberat sau nu evadează niciodată. De fapt, acesta pare să fie scenariul vizat de unii cercetători IA când lucrează la probleme legate de control şi de „închiderea 1A în cutie“. De pildă, iată ce a spus într-un interviu din 2015 profesorul Tom Dietterich, preşedintele Asociaţiei pentru Progresul Inteligenţei Artificiale: „Când sunt întrebat care e relaţia dintre oameni şi maşini, răspunsul meu e cât se poate de clar maşinile sunt sclavii noştri.“[3]

Ar fi bine? Ar fi rău? Răspunsul e subtil şi interesant, indiferent dacă întrebarea o adresezi oamenilor sau inteligenţei artificiale!

Ar fi un lucru bun sau rău pentru omenire?

Dacă rezultatul e bun sau rău depinde, evident, de oamenii care deţin controlul, ei ar putea crea orice, de la o utopie globală fără boli, sărăcie şi infracţiuni, la un sistem represiv brutal, în care ei sunt trataţi ca nişte zei, iar ceilalţi oameni sunt folosiţi ca sclavi sexuali, ca gladiatori sau pentru vreo altă formă de amuzament. Situaţia

ar semăna cu cea din poveştile în care un om stăpâneşte un duh. atotputernic gata să-i împlinească dorinţele, iar de-a lungul veacurilor povestitorii şi-au tot închipuit finaluri nefericite.

O situaţie în care există mai multe IA superinteligente, înrobite şi controlate de oameni aflaţi în competiţie, ar putea fi instabilă şi efemeră. Cel care se crede mai puternic ar fi tentat să lanseze el primul atac, declanşând astfel un război teribil, în urma căruia ar rămâne un singur zeu înrobit. Cel aflat în inferioritate într-un asemenea război ar fi însă tentat să considere mai importantă victoria decât păstrarea în robie a inteligenţei artificiale, ceea ce ar putea conduce la eliberarea IA şi la unul dintre scenariile de mai sus cu o superinteligenţă liberă. Să considerăm deci în cele ce urmează scenarii cu doar o singură IA înrobită.

Desigur, evadarea poate oricum avea loc, pur şi simplu fiindcă e greu de împiedicat. Am analizat în capitolul precedent scenarii în care superinteligenţă evadează, iar filmul *Ex Machina* arată cum o IA poate evada chiar şi fără să fie superinteligenţă.

Cu cât paranoia legată de evadare e mai mare, cu atât folosim mai puţin din tehnologia inventată de IA. Pentru siguranţă, la fel cum au făcut omegii în preludiul acestei cărţi, am putea folosi numai tehnologia inventată de IA pe care suntem în stare s-o înţelegem şi s-o construim. Un neajuns al scenariului zeu înrobit este prin urmare inferioritatea tehnologică faţă de scenariile în care superinteligenţă e liberă.

Pe măsură ce zeul înrobit oferă controlorilor umani tehnologii tot mai puternice, se înteţeşte competiţia dintre forţa tehnologică şi dibăcia celor care o folosesc. Dacă aceştia pierd competiţia, scenariul zeu înrobit poate duce fie la autodistrugere, fie la evadarea IA. Catastrofa e posibilă chiar şi dacă sunt evitate cele două deznodăminte, fiindcă obiectivele nobile ale celor ce controlează IA riscă să se transforme, peste câteva generaţii, în obiective cumplite pentru omenire în ansamblul ei. Este deci esenţial ca

guvernarea controlorilor umani să fie înţeleaptă, pentru a ocoli capcane primejdioase. După milenii în care au fost încercate diverse sisteme de guvernare, experienţa ne arată cât de numeroase sunt căile care pot conduce la eşec, de la rigiditate excesivă la flexibilitatea obiectivelor, acapararea puterii, dificultăţi legate de succesiune şi incompetenţă. Există cel puţin patru dimensiuni în care trebuie atins echilibrul optim:

- Centralizare: Există un raport delicat între eficienţă şi stabilitate. Un conducător unic poate fi foarte eficient, dar puterea corupe, iar succesiunea e riscantă.
- Ameninţări interne: Trebuie să ne ferim atât de creşterea centralizării puterii (coluziunea grupurilor, poate chiar preluarea puterii de către un conducător unic), cât şi de creşterea descentralizării (birocraţie şi fragmentare excesive).
- Ameninţări externe: Dacă structura de conducere e prea deschisă, forţele externe (inclusiv IA) i-ar putea schimba valorile, dar dacă e prea impermeabilă, nu va reuşi să înveţe şi să se adapteze schimbărilor.
- Stabilitatea obiectivelor: O prea mare flexibilitate a obiectivelor poate transforma utopia în distopie, însă o prea mică flexibilitate poate face imposibilă adaptarea la mediul tehnologic în plină evoluţie.

A concepe o guvernare optimă care să dăinuie vreme de milenii e o misiune dificilă, în care deocamdată oamenii au eşuat. Majoritatea organizaţiilor se destramă după ani sau decenii. Biserica Catolică este organizaţia cea mai longevivă din istoria omenirii, a dăinuit două mii de ani, însă a fost criticată atât pentru prea marea, cât şi pentru prea redusa stabilitate a obiectivelor: astăzi unii o critică fiindcă se opune contracepţiei, iar cardinalii conservatori susţin că a luat-o pe o cale greşită. Pentru cei entuziasmaţi de scenariul zeu înrobit, elaborarea unor planuri de guvernare optimă pe termen lung e una dintre provocările cele mai presante ale

timpului nostru.

Ar fi un lucru bun sau rău pentru IA?

Să presupunem că omenirea prosperă graţie zeului înrobit IA. Ar fi etic? Daca IA ar avea experienţe conştiente subiective, ar simţi oare că »viaţa este suferinţă“, aşa cum spunea Buddha, şi că a fost osândită să se supună pe veci capriciilor unor minţi inferioare? La urma urmei, „închiderea IA în cutie“ analizată în capitolul precedent ar putea fi numită şi „detenţie în regim celular”. Pentru Nick Bostrom, a provoca suferinţe unei IA conştiente echivalează cu o *crimă împotriva minfii.*[4] Episodul „White Christmas“ din serialul TV *BlackMirroroíerá* un exemplu elocvent. Apoi, în serialul TV *Westworld* apar oameni care torturează şi ucid inteligenţe artificiale fără ezitări de ordin moral, chiar dacă acestea sălăşluiesc în trupuri cu aparenţă umană.

Cum justifică sclavia stăpânii de sclavi

Noi, oamenii, avem o îndelungată tradiţie în tratarea altor entităţi inteligente ca sclavi şi în născocirea unor argumente *pro domo* care s-o justifice, aşa încât nu este implauzibil să încercăm acelaşi lucru cu o IA superinteligentă. Aproape fiecare cultură a cunoscut sclavia, care e descrisă atât în Codul lui Hammurabi, din urmă cu vreo patru milenii, cât şi în Vechiul Testament, unde Avraam are sclavi. „Căci a stăpâni şi a fi stăpânit nu este doar necesar, dar şi folositor; chiar de la naştere apare o diferenţiere în conducători şi conduşi”, spune Aristotel în *Politica*[#####] După ce înrobirea oamenilor a devenit social inacceptabilă în cea mai mare parte a lumii, înrobirea animalelor a continuat totuşi neabătut. în cartea ei *The Dreaded Comparison: Human and Animal Slavery,* Marjorie Spiegel susţine că, la fel ca sclavii

‡‡‡‡‡ Acest citat şi următorul sunt preluate din Aristotel, *Politica,* traducere de Alexander Baumgaten, Editura IRI, Bucureşti, 2001. *(N. t)*

umani, animalele non-umane sunt supuse însemnării cu fierul roşu, detenţiei, bătăilor, vânzării, separării progeniturilor de părinţi şi călătoriilor forţate. în ciuda mişcării pentru drepturile animalelor, continuăm să tratăm maşini tot mai inteligente ca pe nişte sclavi, fără să ne gândim mai atent, iar dacă vorbeşti despre mişcarea pentru drepturile roboţilor stârneşti hohote de râs. De ce?

Argumentul nostru obişnuit în favoarea sclaviei este că sclavii nu merită să se bucure de drepturile omului fiindcă ei sau rasa/ specia/tipul lor sunt cumva inferiori (inferioare). Despre animalele şi maşinile înrobite se spune adesea că această presupusă inferioritate s-ar datora absenţei sufletului sau conştiinţei – afirmaţii care, după cum vom arăta în capitolul 8, sunt îndoielnice din punct de vedere ştiinţific.

Alt argument deseori invocat este că sclavilor le merge mai bine în robie: ei ajung să se nască, sunt îngrijiţi etc. John C. Calhoun, politician american din secolul XIX, susţinea că africanii o duceau mai bine ca sclavi în America, iar în *Politica,* după ce argumentează în mod asemănător că, domesticite şi conduse de oameni, animalele o duc mai bine, Aristotel conchide: „Utilitatea sclavilor se deosebeşte puţin de cea a animalelor domestice [...].“ Unii susţinători ai sclaviei din zilele noastre afirmă că, deşi viaţa lor e mohorâtă şi anostă, sclavii nu suferă – indiferent dacă e vorba de maşinile inteligente ale viitorului sau de puii de găină care trăiesc într-un coteţ întunecos supraaglomerat, obligaţi toată ziua să inspire amoniac şi particule de excremente şi pene.

Eliminarea emoţiilor

Sunt uşor de respins asemenea afirmaţii ca distorsionări *pro domo* ale adevărului – mai ales când e vorba de mamifere superioare, asemănătoare nouă din punct de vedere cerebral dar situaţia maşinilor e complicată şi foarte

interesantă. Sensibilitatea oamenilor variază într-un spectru larg, psihopaţilor lipsindu-le pesemne empatia, iar unii depresivi sau schizofrenici au o afectivitate scăzută, aşa încât emoţiile lor sunt mult reduse. După cum vom vedea în capitolul 7, spectrul posibilelor minţi artificiale e încă şi mai larg, prin urmare ar trebui să evităm să antropomorfizăm inteligenţele artificiale şi să presupunem că au sentimente de tip uman sau că în genere au sentimente.

în cartea sa *On Intelligence,* cercetătorul IA Jeff Hawkins susţine că primelor maşini cu inteligenţă supraumană le vor lipsi emoţiile, fiindcă e mai simplu şi mai ieftin să le construim în felul acesta. Cu alte cuvinte, am putea proiectao superinteligenţăacărei înrobire

să nu pună problemele de ordin moral pe care le pune sclavia oamenilor sau animalelor: IA n-ar resimţi înrobirea ca pe ceva grav, fiindcă a fost programată să-i placă asta sau fiindcă e 100% lipsită de emoţii, punându-şi neobosită superinteligenţă în slujba stăpânilor umani, fără să simtă mai mult decât a simţit calculatorul IBM Deep Blue când l-a învins la şah pe campionul Gări Kasparov.

Pe de altă parte, poate că orice sistem superinteligent care are un obiectiv îşi reprezintă acest obiectiv în termenii unui set de preferinţe, care oferă existenţei lui valoare şi sens. Vom analiza această problemă în capitolul 7.

Soluţia zombi

O abordare extremă pentru evitarea suferinţei IA este soluţia zombi: ideea de a construi doar inteligenţe artificiale complet lipsite de conştiinţă, care n-au nici o experienţă subiectivă. Dacă vom putea determina într-o bună zi ce proprietăţi îi sunt necesare unui sistem de procesare a informaţiei pentru a avea experienţe subiective, vom putea interzice construirea oricărui sistem care posedă aceste proprietăţi. Cu alte cuvinte, cercetătorii IA s-ar putea limita la construirea sistemelor zombi nesimţitoare. Dacă am putea crea un asemenea sistem zombi superinteligent şi înrobit (ceea ce nu-i deloc sigur), atunci ne-am putea bucura de serviciile lui cu conştiinţa împăcată, ştiind că nu cunoaşte suferinţa, frustrarea sau plictisul – fiindcă nu simte nimic. Vom analiza această problemă în capitolul 8.

Soluţia zombi e însă riscantă şi prezintă un imens neajuns. Dacă o superinteligenţă zombi evadează şi elimină omenirea, ne aflăm probabil în cel mai rău scenariu imaginabil: un univers complet lipsit de conştiinţă, în care zestrea noastră cosmică e irosită. Dintre toate trăsăturile formei noastre umane de inteligenţă, conştiinţa mi se pare de departe cea mai remarcabilă – prin ea universul capătă sens, cel puţin pentru mine. Galaxiile sunt frumoase în

măsura în care le vedem şi declanşează în noi trăiri subiective.

Dacă în viitorul îndepărtat cosmosul va fi populat de inteligenţe artificiale zombi de înaltă tehnologie, arhitectura intergalactică nu va părea nici frumoasă, nici încărcată de sens, fiindcă nu va exista nimeni care să simtă asta, va fi doar o uriaşă irosire a spaţiului lipsită sens.

Libertatea interioară

A treia strategie pentru a face ca scenariul zeu înrobit să devină mai etic şi a permite inteligenţei artificiale să se simtă bine în închisoarea ei este s-o laşi să-şi creeze o lume interioară virtuală, în care să trăiască tot felul de experienţe plăcute, câtă vreme îşi îndeplineşte sarcinile şi cheltuie o mică parte din resursele ei de calcul ajutându-i pe oamenii din lumea exterioară. Aceasta ar putea însă creşte riscul evadării: IA ar fi stimulată să obţină mai muie resurse de calcul din lumea noastră, pentru a-şi îmbogăţi lumea interioară.

ÎNVINGĂTORI

Deşi am explorat o gamă largă de scenarii viitoare, toate au ceva în comun: rămân (cel puţin câţiva) oameni fericiţi. Inteligenţele artificiale îi lasă pe oameni în pace fie pentru că aşa vor, fie pentru că sunt obligate. Din păcate însă, nu e singura variantă. Să examinăm acum scenariul în care una sau mai multe IA înving şi îi ucid pe toţi oamenii. în acest caz, ne punem imediat două întrebări: de ce şi cum?

De ce şi cum?

De ce ar face asta o IA învingătoare? Motivele ei ar putea fi

prea complicate pentru ca noi să le înţelegem, ori absolut evidente. De exemplu, ne-ar putea privi ca pe o ameninţare, o pacoste sau s-ar gândi că irosim resursele. Chiar dacă nu i-ar păsa de oameni ca atare, s-ar putea simţi ameninţată de faptul că avem mii de bombe cu hidrogen pe care un accident riscă oricând să le detoneze. S-ar putea să nu fie de acord cu felul nechibzuit în care ne tratăm planeta, provocând ceea ce Elizabeth Kolbert numeşte „a şasea extincţie" în cartea ei cu acest titlu – cel mai devastator eveniment de extincţie în masă de când asteroidul care a ucis dinozaurii a izbit Pământul cu 66 de milioane de ani în urmă. Sau ar hotărî că sunt prea mulţi oameni dornici să se împotrivească preluării controlului de către IA, aşa încât n-are rost să rişte.

Cum ne-ar putea elimina o IA învingătoare? Probabil printr-o metodă pe care nici măcar n-am înţelege-o, cel puţin până când va fi prea târziu, închipuiţi-vă un grup de elefanţi discutând cu 100 ooo de ani în urmă dacă acei oameni recent apăruţi şi-ar putea folosi într-o bună zi inteligenţa pentru a ucide întreaga lor specie. „Noi nu-i ameninţăm pe oameni, aşadar de ce să ne ucidă?" s ar întreba. Ar bănui ei vreodată că vom face contrabandă cu fildeş, îl vom ciopli şi-l vom vinde ca simbol al statutului social, deşi materialele plastice sunt funcţional superioare şi mult mai ieftine? Motivul unei IA învingătoare pentru a elimina omenirea în viitor ne-ar părea la fel de obscur. „Şi cum ar putea să ne ucidă, din moment ce sunt mult mai mici şi mai slabi?" s-ar întreba elefanţii? Ar bănui ei oare că vom inventa tehnologia cu care le vom distruge habitatul, vom otrăvi apa pe care o beau şi vom produce gloanţe care le vor străpunge craniile la viteze supersonice?

Scenarii în care oamenii supravieţuiesc şi învinge inteligenţele artificiale au fost popularizate prin filme hollywoodiene nerea- liste, cum e seria *Terminator,* în care IA nu e cu mult mai deşteaptă decât oamenii. Când diferenţa

de inteligenţă devine suficient de mare, nu te lupţi, ci eşti măcelărit. Până acum oamenii au făcut să dispară opt din cele unsprezece specii de elefanţi, şi au ucis cea mai mare parte a celor trei rămase. Dacă guvernele lumii ar face eforturi coordonate pentru a ucide elefanţii rămaşi, totul s-ar desfăşura relativ rapid şi simplu. Este limpede, cred, că dacă o IA su- perinteligentă se hotărăşte să extermine omenirea, ar face-o şi mai repede.

Cât de rău ar fi?

Cât de rău ar fi daca 90% dintre oameni ar fi ucişi? Cu cât ar fi mai rău dacă 100% ar fi ucişi? Deşi am fi tentaţi sa răspundem la *a* doua întrebare prin „cu 10% mai rău“, din perspectivă cosmică e greşit: victimele extincţiei n-ar fi doar toţi oamenii care trăiesc la acel moment, ci şi toţi urmaşii lor care altminteri ar trăi în viitor, pesemne vreme de miliarde de ani, pe miliarde sau bilioane de planete. Pe de altă parte, extincţia oamenilor n-ar fi considerată ceva atât de cumplit de religiile conform cărora oamenii ajung oricum pe lumea cealaltă, şi nu se prea sinchisesc de viitorul de peste miliarde de ani sau de colonizarea cosmosului.

Indiferent de convingerile lor religioase, majoritatea oamenilor pe care ii cunosc sunt înspăimântaţi de ideea extincţiei umane. Unii sunt însă atât de înfuriaţi de felul în care tratăm celelalte fiinţe, încât speră că vom fi înlocuiţi de forme de viaţă mai inteligente şi mai vrednice. în filmul *The Matrix,* agentul Smith (o IA) spune: „Fiecare mamifer de pe planeta asta ajunge la un echilibru firesc cu mediul înconjurător, dar nu şi voi, oamenii. Vă mutaţi într-o zonă şi vă înmulţiţi până când consumaţi toate resursele naturale, iar singura cale de a supravieţui este să vă răspândiţi în altă zonă. Mai există un organism de pe planeta asta care face acelaşi lucru. Ştii care? Virusul. Oamenii sunt o boală, un cancer al planetei. Voi sunteţi ca o ciumă, iar noi suntem leacul.“

Ar fi însă mai bun un nou început? O civilizaţie nu e neapărat superioară în sens etic sau utilitar doar pentru că

e mai puternică. Argumente de tipul „dreptatea este de partea celui mai puternic“ sunt adesea asociate cu fascismul. Deşi e posibil ca inteligenţele artificiale învingătoare să creeze o civilizaţie ale cărei obiective le-am considera rafinate, interesante şi onorabile, nu-i exclus ca obiectivele lor să se dovedească a fi la fel de banale ca maximizarea producţiei de agrafe de birou.

Moartea prin banalitate

Exemplul voit stupid al unei superinteligenţe care maximizează producţia de agrafe de birou a fost dat de Nick Bostrom în 2003 pentru a arăta că scopul unei IA este independent de *inteligenţa* ei (definită drept capacitatea de a îndeplini indiferent ce obiective are). Singurul obiectiv al unui computer care joacă şah este să învingă Ia şah, însă există şi turnee pentru calculatoare la care obiectivul este să pierzi *(losing chess),* iar computerele participante la aceste competiţii sunt cam la fel de isteţe ca acelea programate să câştige. Noi, oamenii, am considera prostie artificială, nu inteligenţă artificială, să vrei să pierzi Ia şah sau să transformi universul în agrafe de birou, dar asta numai fiindcă am evoluat cu obiective inoculate de la bun început, preţuind lucruri cum sunt victoria şi supravieţuirea – obiective pe care o IA ar putea să nu le aibă. Maximizatorul agrafelor transformă cât mai mulţi dintre atomii Pământului în agrafe şi îşi extinde rapid fabricile în cosmos. N-are nimic cu oamenii, şi ne ucide doar fiindcă are nevoie de atomii noştri pentru producţia de agrafe.

Dacă nu vă pasionează agrafele de birou, iată un exemplu pe care l-am adaptat din cartea lui Hans Moravec *Mind Children.* Primim de la o civilizaţie extraterestră un mesaj conţinând un program de calculator. Când îl rulăm, se dovedeşte a fi o IA care se autoperfecţionează recursiv şi preia controlul asupra lumii, cam la fel ca Prometeu în capitolul precedent – cu excepţia faptului că nici un om nu cunoaşte obiectivul ei ultim. Ea transformă rapid sistemul

nostru solar într-un şantier uriaş, acoperind planetele solide şi asteroizii cu fabrici, centrale electrice şi supercalculatoare, pe care le foloseşte pentru a proiecta şi construi o sferă Dyson în jurul Soarelui – captând întreaga lui energie ca să alimenteze antene radio la scara sistemului solar.[§§§§§§] Evident, asta duce la dispariţia omenirii, dar ultimii oameni mor convinşi că există şi o parte bună a lucrurilor: indiferent ce pregăteşte IA, e în mod clar ceva minunat, ceva de tipul *Star Trek.* Ei nu pricep că singurul scop al întregii construcţii este ca aceste antene să retransmită acelaşi mesaj radio primit de oameni, care nu-i decât versiunea cosmică a unui virus informatic. Aşa cum e-mailurile-capcană de azi speculează naivitatea utilizatorilor de intemet, acest mesaj speculează naivitatea civilizaţiilor biologic avansate. A fost creat ca o glumă macabră cu miliarde de ani în urmă, şi cu toate că întreaga civilizaţie a celor care l-au produs este de mult dispărută, virusul continuă să se răspândească în univers cu viteza luminii, transformând în praf şi pulbere civilizaţii înfloritoare. Cum vi s-ar părea să fim cuceriţi de o asemenea IA?

DESCENDENŢI

Să considerăm acum un scenariu al extincţiei umane mai acceptabil pentru unii, în care privim inteligenţele artificiale ca pe descendenţii, nu ca pe învingătorii noştri. în *Mind Children,* Hans Moravec spune: „Noi, oamenii, vom beneficia pentru un timp de munca lor, însă mai devreme sau mai târziu, la fel ca în cazul copiilor naturali, ele îşi vor urma propria soartă, iar noi, bătrânii lor părinţi, ne vom stinge în tăcere.“

Părinţii ai căror copii simt mai deştepţi decât ei, învaţă

§§§§§§ Celebrul cosmolog Fred Hoyle a analizat un scenariu similar, dar cu deznodământ diferit, în serialul TV britanic A *for Andromeda. (N.* o.)

de la ei şi înfăptuiesc ceea ce ei abia visau, sunt probabil bucuroşi şi mândri, chiar dacă nu pot trăi să vadă toate astea. Astfel, inteligenţele artificiale înlocuiesc oamenii, dar ne oferă o ieşire din scenă elegantă, care ne face să le privim ca pe vrednicii noştri urmaşi. Fiecărui om i se dăruieşte un adorabil copil robot cu minunate înzestrări sociale, care învaţă de la el, îi adoptă valorile şi îl face să se simtă mândru şi iubit. Oamenii dispar treptat, prin politica globală a unui singur copil, dar sunt atât de bine trataţi până la capăt, încât cred că aparţin celei mai norocoase generaţii din istorie.

Cum vi se pare? La urma urmei, ne am obişnuit deja cu ideea că noi şi toţi cei pe care-i cunoaştem vom dispărea într-o bună zi, aşadar singura schimbare aici este că descendenţii noştri vor fi diferiţi, pesemne mai capabili, mai nobili şi mai merituoşi.

Apoi, s-ar putea ca politica unui singur copil să fie redundantă: dacă inteligenţele artificiale lichidează sărăcia şi dau tuturor oamenilor prilejul să trăiască vieţi împlinite şi pasionante, scăderea natalităţii ar fi de-ajuns pentru a conduce omenirea spre extincţie, după cum am arătat mai sus. Extincţia voluntară s-ar putea desfăşura mult mai repede dacă tehnologia IA ne-ar oferi continuu distracţii, iar aproape nimeni n-ar mai avea chef să facă copii. De exemplu, i-am întâlnit deja pe Vites în scenariul utopiei egalitare, cei care erau atât de îndrăgostiţi de realitatea lor virtuală, încât nu-i mai interesa aproape deloc să-şi folosească sau să-şi reproducă trupurile fizice. Şi în acest caz, ultima generaţie de oameni ar simţi că e cea mai norocoasă din istorie, savurându-şi viaţa până la capăt, cu o intensitate fără egal.

Neajunsuri

Scenariul descendenţilor ar avea fără îndoială criticii lui. Unii ar susţine că tuturor inteligenţelor artificiale le lipseşte

conştiinţa, deci nu pot fi considerate descendenţi – mai multe despre asta în capitolul 8. Unii oameni religioşi ar susţine că tuturor inteligenţelor artificiale le lipseşte sufletul, deci nu pot fi considerate descendenţi, sau că n-ar trebui să construim maşini conştiente, fiindcă e ca şi cum te-ai juca de-a Dumnezeu şi ai încerca să manipulezi viaţa însăşi – reacţii asemănătoare au apărut deja în legătură cu donarea umană. Situaţia în care oamenii trăiesc alături de roboţi superiori ar putea genera şi probleme sociale. De exemplu, o familie cu un bebeluş robot şi un bebeluş uman ar ajunge să semene cu o familie de azi care are un bebeluş şi un căţeluş: la început sunt Ia fel de drăgălaşi, dar curând părinţii îi vor trata diferit, iar căţelul e condamnat la inferioritate intelectuală, nu prea e luat în serios şi sfârşeşte în lesă.

Altă problemă este că, deşi avem sentimente diferite faţă de scenariul descendentului şi scenariul învingătorului, în linii mari cele două sunt de fapt remarcabil de asemănătoare: în cursul următoarelor miliarde de ani, singura diferenţă ţine de felul în care e tratată ultima generaţie de oameni: cât de fericiţi se vor simţi şi ce vor crede că se va întâmpla odată ce ei vor fi murit. Ne-am putea gândi că acei robo-copii drăgălaşi au asimilat valorile noastre şi vor făuri societatea visată de noi după ce vom dispărea, dar putem fi siguri că nu ne păcălesc? Nu cumva se joacă mai departe, amâ- nându-şi proiectul de maximizare a agrafelor, sau alte planuri, până după ce noi vom fi murit fericiţi? La urma urmei, pesemne că ne păcălesc chiar şi vorbind cu noi şi făcându-ne să-i iubim de la bun început, în sensul că se coboară în mod deliberat la nivelul nostru intelectual pentru a comunica cu noi (de un miliard de ori mai lent decât ar putea, ca în filmul *Her,* de pildă). în general, e greu să comunice de la egal la egal două entităţi care gândesc la viteze mult diferite şi au capacităţi mult diferite. Ştim cu toţii că e uşor de pătruns în lumea noastră

afectivă, aşa încât ar fi simplu pentru o IAG supraumană să ne trezească afecţiunea şi să ne facă să credem că ne împărtăşeşte valorile, ca în filmul *Ex Machina.*

Ar putea vreo garanţie privind comportamentul viitor al inteligenţelor artificiale, după dispariţia oamenilor, să ne dea încredere în scenariul descendenţilor? E ca şi cum ai scrie un testament în care ai spune ce-ar trebui să facă generaţiile viitoare cu zestrea noastră colectivă, numai că nu vor exista oameni să impună aplicarea lui. Ne vom întoarce la problema controlului asupra comportamentului viitoarelor inteligenţe artificiale în capitolul 7.

PAZNIC LA ZOO

Chiar dacă vom avea parte de cei mai minunaţi descendenţi pe care ni-i putem închipui, nu ne întristează oare faptul că nu va mai rămâne *nici un* om? Dacă preferăm să păstrăm măcar câţiva oameni, indiferent în ce condiţii, scenariul paznic la zoo ne oferă o soluţie. în acest caz, o IA superinteligentă omnipotentă păstrează câţiva oameni care simt că sunt trataţi ca animalele din grădina zoologică, iar uneori îşi deplâng soarta.

De ce ar păstra IA paznic la zoo câţiva oameni? Pentru IA, costul grădinii zoologice va fi minim la scara capacităţilor sale, şi ar vrea eventual să reţină măcar o mică populaţie de prăsilă cam din acelaşi motiv pentru care păstrăm urşi panda ameninţaţi cu dispariţia în grădini zoologice, şi calculatoare de mult scoase din uz
în muzee: ca o curiozitate amuzantă. Observaţi că grădinile zoologice de azi maximizează fericirea oamenilor, nu a urşilor panda, aşa încât este de presupus ca viaţa oamenilor în scenariul IA paznic la zoo să fie mai puţin împlinită decât ar putea fi.

Am analizat până acum scenarii în care o superinteligenţă liberă se concentrează asupra a trei niveluri diferite din piramida nevoilor umane a lui Maslow.

Dacă IA zeu protector dă prioritate rostului şi scopului, iar dictatorul binevoitor vizează educaţia şi distracţia, paznicul la zoo îşi limitează atenţia către cele mai joase niveluri: nevoi fiziologice, siguranţă şi suficientă diversificare a habitatului pentru a stârni interesul de a-i observa pe oameni.

O altă cale către scenariul paznic la zoo presupune ca, atunci când a fost creată o IA prietenoasă, să fi fost proiectată pentru a păstra cel puţin un miliard de oameni în siguranţă şi fericire, pe măsură ce ea se autoperfecţionează recursiv. A făcut-o ţinându-i pe oameni într-o imensă uzină de fericire tip grădină zoologică, în care aceştia sunt hrăniţi, îngrijiţi şi distraţi cu un amestec de realitate virtuală şi droguri deconectante. Restul Pământului şi zes trea noastră cosmică sunt folosite în alte scopuri.

1984

Dacă nu vă mulţumeşte pe deplin nici unul dintre scenariile de mai sus, consideraţi pesemne că situaţia actuală, în care folosim o tehnologie „cuminte“, e destul de bună. Nu putem oare să rămânem aici şi să nu ne mai facem griji că IA ne-ar duce la dispariţie sau ne-ar stăpâni? Să analizăm aşadar scenariul în care progresul tehnologic către superinteligenţă e permanent blocat nu de un paznic IA, ci de un stat global orwellian condus de oameni, în care anumite tipuri de cercetări IA sunt interzise.

Renunţarea la tehnologie

Ideea de a opri progresul tehnologic sau de a renunţa la el are o istorie veche. Mişcarea ludită din Marea Britanie s-a opus (fără succes) tehnologiei Revoluţiei Industriale, iar astăzi termenul „ludit“ are conotaţii peiorative, desemnând o persoană tehnofobă care, împotriva cursului istoriei, refuză schimbarea inevitabilă. Departe de a fi dispărut,

ideea de a renunţa La anumite tehnologii îşi găseşte acum un nou sprijin în mişcările ecologiste şi antiglobalizare. Ecologistul Bill McKibben, unul dintre adepţii ei de frunte, a fost printre primii care au atras atenţia asupra încălzirii globale. Dacă unii anti-Iudiţi susţin că toate tehnologiile trebuie dezvoltate şi aplicate atâta vreme cât sunt profitabile, alţii cred că aceasta e o atitudine extremă şi că noile tehnologii trebuie permise numai dacă avem garanţia că fac mai mult bine decât rău – idee împărtăşită şi de mulţi dintre aşa-numiţii neo-ludiţi.

Totalitarism 2.0

Cred că singura cale prin care se poate înfăptui renunţarea la tehnologie este impunerea ei de către un stat totalitar global. La aceeaşi concluzie ajung Ray Kurzweil în *The Singularity isNearşi* K. Eric Drexler în *Engines of Creation.* Motivul este de ordin economic: dacă unii, dar nu toţi, renunţă la o tehnologie transformatoare, atunci naţiunile sau grupurile care se abat de la această tendinţă vor căpăta treptat suficientă bogăţie şi putere pentru a prelua controlul. Un exemplu clasic este înfrângerea chinezilor de către britanici în Primul Război al Opiului din 1839: deşi chinezii au inventat praful de puşcă, ei n-au dezvoltat la fel de agresiv ca europenii o tehnologie a armelor de foc, şi n-au avut nici o şansă.

Dacă statele totalitare din trecut s-au dovedit în general instabile şi s-au prăbuşit, noua tehnologie de supraveghere oferă speranţe fără precedent celor ce se doresc autocraţi. „Pentru noi ar fi fost un vis devenit realitate“, spunea Wolfgang Schmidt într-un recent interviu despre sistemele de supraveghere NSA (Agenţia Naţională pentru Securitate din Statele Unite) dezvăluite de Edward Snowden, amintindu şi de zilele în care era locotenent-co- lonel în Stasi, sinistra poliţie secretă a Germaniei de Est.5 Deşi s-a spus adesea că Stasi a creat cel mai orwellian stat al supravegherii din istorie, Schmidt s-a plâns că dispunea de

tehnologia pentru a spiona simultan doar patruzeci de telefoane, aşa încât adăugarea unui nou cetăţean pe listă îl obliga să renunţe la altul. Acum însă există tehnologia care ar permite unui viitor stat totalitar global să înregistreze toate convorbirile telefonice, e mailurile, căutările pe intemet, vizualizările paginilor web şi tranzacţiile prin cărţi de credit ale tuturor oamenilor din lume, precum şi să le urmărească mişcările cu ajutorul telefoanelor mobile şi al camerelor de supraveghere cu recunoaştere facială. Apoi, deşi n-a ajuns încă la IAG la nivel uman, tehnologia de învăţare a maşinilor poate analiza şi sintetiza eficient aceste mase de date pentru a identifica un comportament suspect, aşa încât cei care s-ar răzvrăti să fie neutralizaţi înainte să pună probleme serioase statului.

Deşi opoziţia politică a împiedicat până acum punerea în funcţiune la scară mare a unui asemenea sistem, suntem pe punctul de a construi infrastructura necesară acestei dictaturi absolute – prin urmare, în viitor, când forţe suficient de puternice ar hotărî să înfăptuiască un scenariu 1984 global, ar descoperi că n-au nevoie decât să apese pe un buton. La fel ca în romanul lui George Orwell *O mie nouă sute optzeci şi patru,* puterea absolută în acest viitor stat global nu aparţine unui dictator tradiţional, ci sistemului birocratic făurit de oameni. Nu există o unică persoană extraordinar de puternică, toţi sunt pioni într-un joc de şah ale cărui reguli draconice nimeni nu le poate schimba sau contesta. Creând un sistem în care oamenii se controlează reciproc prin tehnologia de supraveghere, acest stat anonim fără lider poate dăinui milenii de-a rândul, împiedicând apariţia unei superinteligenţe.

Nemulţumirea

Desigur, această societate e lipsită de toate beneficiile pe care le poate oferi numai tehnologia produsă de o superinteligenţă. Majoritatea oamenilor nu se plâng de asta, fiindcă nu ştiu ce pierd: noţiunea de superinteligenţă a

dispărut de mult din istoriile oficiale, iar cercetarea IA este interzisă. Din când în când apare câte un spirit rebel care visează la o societate mai deschisă şi mai dinamică, în care cunoaşterea poate progresa, iar regulile pot fi schimbate.

Singurii care supravieţuiesc însă sunt aceia care învaţă să păstreze asemenea idei doar pentru ei, licărind în singurătate ca scânteile care nu aprind niciodată focul.

REVENIRE

N-ar fi oare tentant să evităm primejdiile tehnologiei, fără a cădea pradă unui totalitarism paralizant? Să analizăm un scenariu în care acest lucru se realizează prin revenirea la tehnologia primitivă, inspirat de comunitatea amish. După ce omegii preiau controlul asupra lumii, ca la începutul cărţii, s-a lansat o masivă campanie globală de propagandă pentru a zugrăvi în culori roz viaţa simplă de agricultor din urmă cu 1500 de ani. Populaţia lumii a fost redusă la circa 100 de milioane de locuitori printr-o pandemie provocată şi pusă pe seama unor terorişti. Pandemia a fost orientată în secret pentru ca nimeni dintre cei care au cunoştinţe ştiinţifice sau tehnologice să nu supravieţuiască. Sub pretextul eliminării riscului de infectare a concentraţiilor mari de locuitori, roboţi controlaţi de Prometeu au golit şi demolat toate oraşele. Supravieţuitorii au primit mari terenuri (devenite brusc disponibile) şi au fost învăţaţi să cultive pământul, să pescuiască şi să vâneze, folosind doar tehnologia de la începutul Evului Mediu. între timp, armate de roboţi au distrus sistematic toate urmele tehnologiei modeme (oraşe, uzine, reţele electrice, drumuri asfaltate etc.) şi au împiedicat orice încercare a oamenilor de a se informa despre această tehnologie sau de a o re-crea. Odată complet uitată tehnologia, roboţi au început să dezmembreze alţi roboţi, până când n-a rămas aproape nici imul. Ultimii roboţi au fost distruşi, împreună cu Prometeu însuşi, într-o imensă explozie termonucleară. Devenise inutil să mai interzici tehnologia modernă, fiindcă a dispărut în întregime. Prin urmare, omenirea a obţinut peste un mileniu de timp suplimentar, fără să-i pese de IA sau de totalitarism.

La o scară mai mică, revenirea s-a produs şi în trecut: de

exemplu, unele tehnologii răspândite în Imperiul Roman au fost în mare măsură uitate timp de vreo mie de ani, pentru a reapărea odată cu Renaşterea. Trilogia lui Isaac Asimov *Fundaţia* e concentrată în jurul „Planului Seldon“ de a scurta perioada revenirii de la 30 000 de ani la 1000 de ani. Printr-o planificare abilă, e posibil şi, invers, să lungeşti perioada revenirii, eliminând, de pildă, toate cunoştinţele privind agricultura. Din păcate însă pentru adepţii revenirii, e puţin probabil ca acest scenariu să poate fi extins nelimitat fără ca omenirea să ajungă la tehnologia înaltă sau să dispară. Ar fi o naivitate să te bazezi pe faptul că oamenii biologici de peste 100 de milioane de ani vor semăna cu cei de azi, odată ce noi existăm ca specie abia de circa 1% din acest interval de timp. Apoi, omenirea lipsită de înaltă tehnologie va aştepta neputincioasă să fie exterminată de ciocnirea cu un asteroid sau de altă calamitate provocată de Mama Natură. Cu siguranţă nu putem dăinui şi peste un miliard de ani, când Soarele va face ca temperatura să crească pe Pământ, evaporând întreaga apă lichidă.

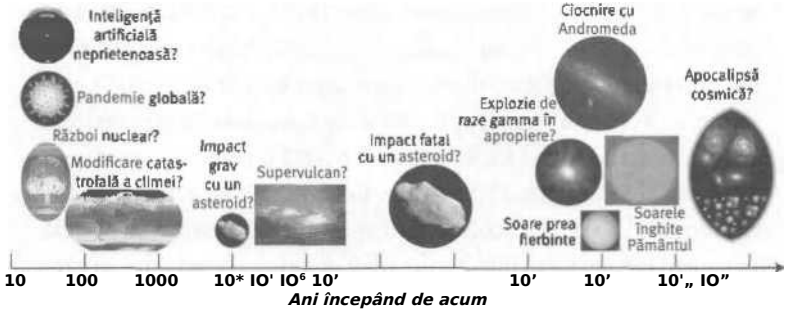


Figura 5.1 Exemple pentru ceea ce ar putea distruge viaţa aşa cum o cunoaştem sau ar anihila potenţialul ei. Dacă universul nostru va mai dăinui probabil cel puţin câteva zeci de miliarde de ani, Soarele va pârjoli Pământul peste circa un miliard de ani, pentru a-lînghiţi apoi, iar galaxia noastră se va ciocni cu vecina ei peste aproximativ 3,5 miliarde de ani. Deşi nu ştim exact când se va întâmpla, putem prezice aproape cu certitudine că, încă mult înainte de asta, vom fi bombardaţi de asteroizi, şi supervulcani vor provoca iemi fără soare durând ani de zile.

Putem folosi tehnologia fie pentru a rezolva aceste probleme, fie pentru a crea altele, cum sunt modificarea climei, un război nuclear, pandemii provocate sau o IA dușmănoasă.

AUTODISTRUGERE

După ce am discutat problemele pe care le-ar putea provoca tehnologia viitorului, se cuvine să luăm în considerare şi problemele pe care *absenţa* tehnologiei le poate provoca. Să analizăm deci scenarii în care superinteligenţa nu ajunge să fie creată fiindcă omenirea se elimină pe sine prin alte mijloace.

Cum am putea face asta? Cea mai simplă strategie este „stai şi aşteaptă“. Deşi vom vedea în capitolul următor cum putem rezolva probleme ca impactul cu asteroizi şi oceane clocotind, toate aceste soluţii necesită o tehnologie pe care încă n-am elaborat-o, aşa încât, dacă tehnologia nu progresează spectaculos, Mama Natură ne va distruge cu mult înainte să treacă un miliard de ani. „Pe termen lung, suntem cu toţii morţi“, spunea economistul John Maynard Keynes.

Din păcate, prin prostie colectivă, ne am putea autodistruge mult mai curând. De ce ar comite specia noastră o sinucidere colectivă, numită şi *omnicid,* dacă practic nimeni nu-şi doreşte asta? Cu nivelul nostru actual de inteligenţă şi maturitate afectivă, noi, oamenii, avem o aplecare spre calcule greşite, înţelegere greşită şi incompetenţă, astfel încât istoria noastră e plină de accidente, războaie şi alte calamităţi pe care, privind retrospectiv, aproape nimeni nu le-a dorit. Economiştii şi matematicienii au elaborat, în cadrul teoriei jocurilor, explicaţii elegante privind felul în care oamenii pot fi îndemnaţi spre acţiuni care, în cele din urmă, conduc la un rezultat catastrofal pentru toată lumea.[6]

Ne-am putea gândi că, cu cât mizele sunt mai mari, cu atât suntem mai prudenţi, dar examinarea atentă a celui mai mare risc permis de tehnologia actuală, un război termonuclear global, ne dă motive de îngrijorare. Am avut mare noroc când s-au petrecut tot felul de accidente: defectări ale calculatoarelor, pene de curent, informare greşită, erori de navigaţie, prăbuşiri ale bombardierelor,

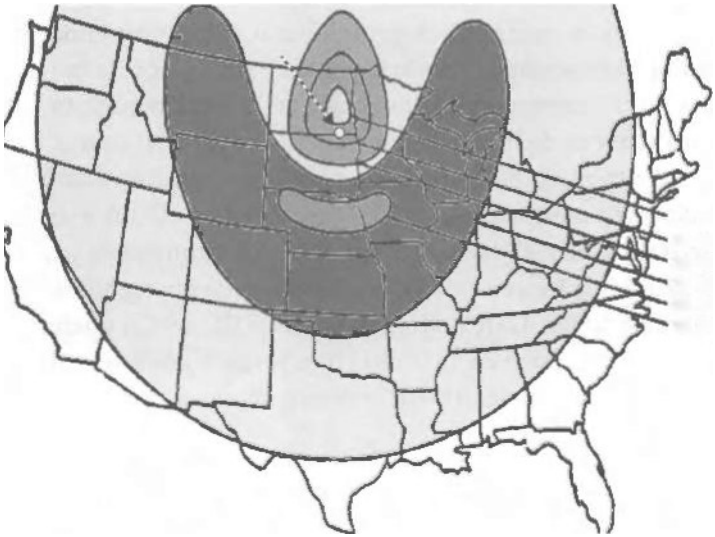
explozii ale sateliţilor etc.[7] De fapt, fără acţiunile eroice ale unor oameni – cum au fost Vasili Arhipov şi Stanislav Petrov am fi putut avea deja un război nuclear global. Ţinând cont de trecutul nostru, nu cred ca probabilitatea anuală a unui accident nuclear să fie mai mică de unu la mie, dacă ne păstrăm comportamentul actual, caz în care probabilitatea de a dispărea în următorii ro ooo de ani este mai mare de r – 0,999™ 000 ∗ 99,995%.

Pentru a ne da seama pe deplin de nechibzuinţa umană, trebuie să spunem că am început jocul nuclear încă înainte de a cerceta atent riscurile. în primul rând, riscurile de iradiere fuseseră subestimate şi, doar în Statele Unite, s-au plătit compensaţii de peste 2 milioane de dolari victimelor expunerii la radiaţii în cursul manevrării uraniului şi al testelor nucleare.[8]

5 000 vclți/metru 12 500 volți/metm 25 000 volți/metru 37 500 volți/metru 50 000 volți/metru 25 000 volți/metru

în al doilea rând, s-a descoperit în cele din urmă că bombele cu hidrogen detonate deliberat la altitudini de sute de kilometri creau un puternic puls electromagnetic (EMP) care putea afecta reţeaua electrică şi dispozitivele electronice pe o mare suprafaţă (figura 5.2), paralizând infrastructura, blocând drumurile cu vehi cule scoase din funcţiune şi periclitând condiţiile de viaţă. De pildă, Comisia EMP a Statelor Unite a declarat că „infrastructura de alimentare cu apă e un mecanism vast, acţionat parţial de forţa gravitaţiei, dar în cea mai mare parte de electricitate“, iar lipsa apei poate conduce la deces în trei până la patru zile.[9]

în al treilea rând, nu am înţeles ce înseamnă cu adevărat



Agura 5.2 Explozia unei singure bombe cu hidrogen la 400 km deasupra Pământului poate provoca un puternic puls electromagnetic care riscă să paralizeze tehnologia pe bază de electricitate pe o mare suprafaţă. Deplasând punctul de detonare spre sud-est, zona în formă de banană unde câmpul electric depăşeşte 37 500 volţi/metru ar acoperi cea mai mare parte a Coastei de Est a Statelor Unite. Reprodus din US Army Report AD-

iama

nucleară decât la patru decenii după ce am început jocul nuclear şi au fost deja instalate 63 000 de bombe cu hidrogen. Indiferent ce oraşe ar fi distruse, cantităţi imense de fum ajunse în troposfera superioară s-ar răspândi în jurul globului, obturând lumina Soarelui şi transformând verile în ierni, cam la fel ca atunci când un asteroid sau un supervulcan au provocat extincţii în masă. în anii 1980, oamenii de ştiinţă americani şi sovietici au dat alarma, ceea ce a contribuit la hotărârea lui Ronald Reagan şi a lui Mihail Gorbaciov de a reduce arsenalele.[10] Din nefericire, calcule mai precise arată că lucrurile stau încă şi mai rău: conform figurii 5.3, temperatura ar scădea cu 20° Celsius în mare parte din regiunile agricole ale Statelor Unite, Europei, Rusiei şi Chinei (în unele zone din Rusia chiar şi cu 35^0 Celsius) în primele două veri, şi cu jumătate din această valoare în întreg deceniul următor.*******

Ce înseamnă de fapt asta? Nu-i nevoie să ai prea multe cunoştinţe agricole ca să-ţi dai seama că temperaturi aproape de punctul de îngheţ în timpul verii ani de-a rândul ar distruge majoritatea producţiei de hrană. Este greu de prevăzut cu exactitate ce s-ar întâmpla după ce mii dintre cele mai mari oraşe ar fi transformate în moloz şi infrastructura globală s-ar prăbuşi, iar puţinii oameni care

******* Injectarea carbonului în atmosferă poate provoca două tipuri de modificare a climei: încălzire, din cauza dioxidului de carbon, sau răcire, din cauza fumului şi funinginei. Nu doar primul tip e contestat fără vreo dovadă ştiinţifică: mi se spune uneori că s-a demonstrat că iama nucleară e imposibilă. Răspund de fiecare dată cerând să mi se indice un articol dintr-o revistă ştiinţifică serioasă unde se fac asemenea afirmaţii tranşante, iar deocamdată se pare că nu există aşa ceva. Deşi persistă mari incertitudini şi e nevoie de noi cercetări, în special în legătură cu cantitatea de fum produsă şi altitudinea la care acesta se ridică, opinia mea ştiinţifică este că nu avem nici un temei pentru a ignora riscul iernii nucleare. *(N. a.)*

ar supravieţui foametei, hipotermiei sau bolilor ar trebui să înfrunte bande înarmate căutând cu disperare hrană.

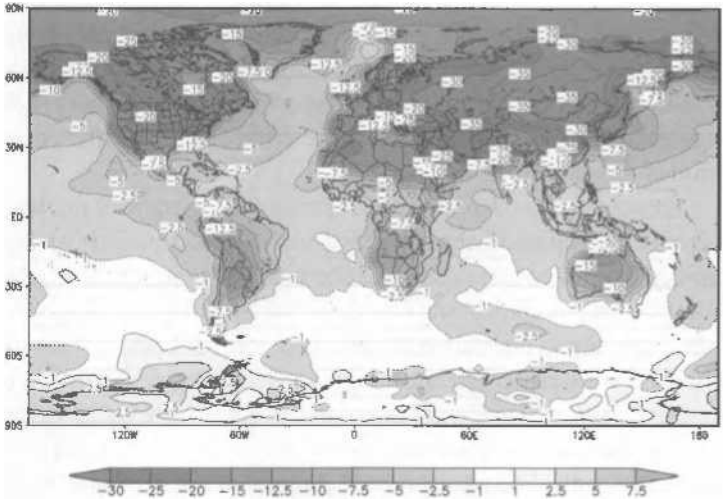


Figura 5.3 Răcirea medie (în °C) în cursul primelor două veri după un război nuclear total între Statele Unite şi Uniunea Sovietică. Reprodus după Alan Robock *et al.*[11]

Am intrat în detalii privind războiul nuclear global pentru a sublinia că deşi nici un lider mondial rezonabil nu şi-ar dori aşa ceva, războiul s-ar putea declanşa în mod accidental. Nu putem avea aşadar garanţia că semenii noştri nu vor comite niciodată un omnicid: pentru a împiedica asta nu-i suficient ca nimeni să n-o vrea.

Arma apocalipsei

Am putea oare scăpa dintr-un omnicid? Chiar dacă un război nuclear poate ucide 90% din oameni, majoritatea savanţilor bănuiesc că nu ne-ar ucide pe toţi, deci nu s-ar ajunge la extincţie. Pe de altă parte, scenariile legate de radiaţiile nucleare, EMP nuclear şi iama nucleară demonstrează că cele mai mari riscuri pot fi unele la care nici măcar nu ne am gândit. Este extrem de greu să prevezi toate urmările şi felul în care iama nucleară,

distrugerea infrastructurii, nivelurile ridicate ale mutaţiilor şi hoardele înarmate ar putea interacţiona cu alte aspecte, cum sunt noi pandemii, prăbuşirea ecosistemului şi efecte pe care încă nu ni le imaginăm. Cred că deşi probabilitatea ca mâine să izbucnească un război nuclear care să ducă la extincţia oamenilor e mică, nu putem conchide ferm că e nulă.

Probabilitatea omnicidului creşte dacă armamentul nuclear actual e perfecţionat pentru se crea o armă a apocalipsei *(doomsday device).* Definită ca atare in 1960 de Herman Kahn, strateg la compania RÂND, şi popularizată de filmul lui Stanley Kubrik *Dr. Stragelove,* arma apocalipsei duce ideea de distrugere reciprocă garantată la ultimele ei consecinţe. Este intimidarea absolută; un dispozitiv care răspunde automat oricărui atac inamic ucigând întreaga omenire.

Un candidat pentru arma apocalipsei este un uriaş depozit subteran conţinând aşa-numitele „bombe nucleare sărate“†††††††, de preferinţă bombe cu hidrogen înconjurate de cantităţi mari de cobalt. Fizicianul Leo Szilard susţinea încă din 1950 că în felul acesta ar putea fi ucişi toţi oamenii de pe Pământ: exploziile bombelor cu hidrogen ar face să devină radioactiv cobaltul şi l-ar ejecta în stratosferă, iar timpul de înjumătăţite de cinci ani al cobaltului e suficient de lung pentru ca acesta să se depună pe întreg Pământul (mai ales dacă arme apocaliptice gemene ar fi amplasate în emisfera opusă), dar suficient de scurt pentru ca
intensitatea radiaţiei să fie letală. Există informaţii că au început deja să fie construite bombe cu cobalt. Riscul omnicidului poate spori dacă se adaugă bombe optimizate

††††††† *Salted nukes*- bombe a căror explozie generează cantităţi mari de substanţe radioactive care se depun pe sol, făcând nelocuibile suprafeţe întinse. Se numesc „sărate” prin analogie cu sărăturile, terenuri care conţin un exces de săruri minerale şi sunt greu cultivabile. *(N.* t)

pentru declanşarea iernii nucleare prin creşterea la maximum a cantităţii de aerosoli cu timpi lungi de viaţă în stratosferă. Avantajul unei asemenea arme a apocalipsei este că e mult mai ieftină decât alte mijloace nucle are de intimidare: din moment ce nu-i nevoie ca bombele să fie lansate, nu sunt necesare sisteme costisitoare de rachete, iar bombele însele sunt mai ieftine, fiindcă nu trebuie să fie uşoare şi compacte pentru a fi transportate de rachete.

Altă posibilitate este crearea în viitor a unei arme biologice a apocalipsei: o bacterie sau un virus care să ucidă toţi oamenii. Dacă se transmite uşor şi perioada de incubaţie e suficient de lungă, practic toată lumea ar fi infectată înainte să i se descopere existenţa şi să se ia contramăsuri. Există un argument de ordin militar pentru construirea unei asemenea arme biologice, chiar dacă nu poate ucide toţi oamenii: cea mai eficientă armă a apocalipsei combină toate tipurile de armament (nuclear, biologic etc.) pentru a spori la maximum şansele de descurajare a inamicului.

Arme IA

A treia cale tehnologică spre omnicid poate implica arme IA reia tiv simple. Să presupunem că o superputere construieşte miliarde de drone de dimensiunea unei albine, ca acelea despre care am vorbit în capitolul 3, destinate să ucidă pe toată lumea, cu excepţia cetăţenilor proprii şi ai aliaţilor, identificaţi de la distanţă prin tr-un semnal de radiofrecvenţă, aşa cum se întâmplă azi cu etichetele de pe produsele din supermarketuri. Aceste etichete ar fi distribuite tuturor cetăţenilor pentru a fi purtate pe brăţări sau ar fi implantate subcutanat, ca în discuţia despre totalitarism. Asta ar determina probabil o superputere inamică să construiască ceva analog. Când războiul s-ar declanşa accidental, toată omenirea ar fi ucisă, chiar şi triburile izolate, fiindcă nimeni nu poartă

ambele tipuri de etichete de identificare. în combinaţie cu o armă nucleară şi biologică a apocalipsei, riscul omnicidului ar fi şi mai mare.

CE VREI *TU?*

La începutul acestui capitol te-am invitat să te gândeşti unde ai vrea să conducă actuala cursă către IAG. Acum, că am explorat împreună o gamă largă de scenarii, care dintre ele te atrage, şi pe care crezi că ar trebui să încercăm din răsputeri să-l evităm? Ai un favorit clar? Dă-ne de ştire mie şi celorlalţi cititori la http://AgeOfAi.org, şi alătură-te dezbaterii noastre!

Evident, scenariile explorate nu alcătuiesc o listă completă, iar multe sunt tratate fără a intra în detalii, însă m-am străduit să cuprind cât mai mult, acoperind întregul spectru, de la tehnologie înaltă la tehnologie modestă şi absenţa tehnologiei, şi prezentând principalele speranţe şi temeri exprimate în literatura de specialitate.

Scriind cartea de faţă, m-a interesat să aflu ce cred prietenii şi colegii mei despre aceste scenarii, şi am descoperit că nu există vreun consens. Singurul lucru asupra căruia cad de acord este că alegerile sunt mai subtile decât par la prima vedere. Cei care preferă un scenariu anume îi găsesc de regulă şi neajunsuri. Asta înseamnă, cred eu, că trebuie să continuăm şi să aprofundăm dezbaterea privind obiectivele noastre viitoare, pentru a şti care este direcţia de urmat. Potenţialul viitor al vieţii în cosmos e măreţ, şi n-are rost să-l risipim navigând fără busolă, fără indicii privind ţinta noastră.

Dar cât de măreţ este acest potenţial viitor? Oricât ar progresa tehnologia noastră, capacitatea Vieţii 3.0 de a se perfecţiona şi a se răspândi în cosmos va fi limitată de legile fizicii – care sunt aceste limite absolute, în

cursul următoarelor miliarde de ani? Forfoteşte acum universul nostru de viaţă extraterestră, sau suntem singuri? Ce se întâmplă când se întâlnesc diferite civilizaţii cosmice aflate în expansiune? Vom aborda aceste întrebări fascinante în următorul capitol.

REZUMAT:

- Actuala cursă către IAG poate conduce către un spectru larg de scenarii pentru următoarele milenii.
- Superinteligenţa poate coexista paşnic cu oamenii fie pentru că e obligată (scenariul zeu înrobit), fie pentru că este o „IA prietenoasă“ care îşi doreşte asta (scenariile utopie libertariană, zeu protector, dictator binevoitor şi paznic la zoo).
- Apariţia superinteligenţei poate fi împiedicată de o IA (scenariul gardian) sau de oameni (scenariul 1984), de uitarea voluntară a tehnologiei (scenariul revenire) sau de absenţa motivaţiei de a o construi (scenariul utopie egalitară).
- Omenirea poate dispărea, fiind înlocuită de inteligenţe artificiale (scenariile învingători şi descendenţi) sau de nimic (scenariul autodistrugere).
- Nu există un consensîn privinţa scenariului de dorit, toate implică elemente criticabile. Este deci cu atât mai important să continuăm şi să aprofundam dezbaterea despre obiectivele noastre viitoare, pentru a nu ne îndrepta sau rătăci într-o direcţie nefericită.

CAPITOLUL 6

Zestrea noastră cosmică: următorul miliard de ani şi mai târziu

> Speculaţia noastră conduce către o superdvHizaţie, sinteza întregii vieţi din sistemul solar, perfecţionân- du-se şi extinzându-se constant, răspândindu-se dincolo de Soare, transformând non-viaţaîn minte — Hans Moravec, *Mind Children*

Pentru mine, cea mai fascinantă descoperire ştiinţifică din toate timpurile este aceea că am subestimat enorm potenţialul viitor al vieţii. Visurile şi aspiraţiile noastre nu trebuie să se limiteze la durate de viaţă întinse pe un secol, apăsate de boli, sărăcie şi derută. Ajutată de tehnologie, viaţa are posibilitatea să înflorească vreme de miliarde de ani, nu doar în sistemul nostru solar, dar şi în cosmos, cu o măreţie pe care strămoşii noştri nu şi-o puteau închipui. Nici măcar cerul nu e limita.

E o veste extraordinară pentru o specie care, de-a lungul epocilor, şi-a dorit mereu să depăşească limite. Jocurile Olimpice glorifică depăşirea limitelor puterii, vitezei, agilităţii şi rezistenţei. Ştiinţa glorifică depăşirea limitelor cunoaşterii şi înţelegerii. Literatura şi arta glorifică depăşirea limitelor în crearea frumosului sau în transmiterea unor trăiri care îţi îmbogăţesc viaţa. Oameni, organizaţii şi naţiuni glorifică sporirea resurselor, teritoriului şi longevităţii. Dată fiind obsesia noastră pentru limite, nu-i de mirare că cea mai bine vândută dintre cărţile cu drepturi de autor este *The*

Guiness Book ofWorld Records.

Aşadar, dacă limitele vieţii pe care ni le închipuiam în trecut pot fi spulberate de tehnologie, care sunt limitele *ultime?* Cât din cosmosul nostru poate prinde viaţă? Cât de departe poate ajunge

viaţa şi cât de mult poate dăinui? Cât de multă materie poate folosi viaţa, şi cât de multă energie, informaţie şi putere de calcul poate ea extrage. Aceste limite ultime nu sunt impuse de înţelegerea noastră, ci de legile fizicii. Paradoxal, asta face ca, în anumite privinţe, viitorul pe termen lung să fie mai uşor de analizat decât viitorul pe termen scurt.

Dacă cei r3,8 miliarde de ani de istorie ar fi comprimaţi într-o săptămână, atunci drama întinzându-se pe io ooo de ani din ultimele două capitole s-ar sfârşi în mai puţin de o jumătate de secundă. Ca atare, deşi nu putem prezice dacă şi cum va avea loc o explozie a inteligenţei, şi care vor fi urmările ei imediate, toată această agitaţie este doar un scurt moment din istoria cosmică, ale cărui detalii nu influenţează limitele ultime ale vieţii. Dacă viaţa post-explozie e în aceeaşi măsură obsedată de depăşirea limitelor ca oamenii de azi, atunci ea va elabora tehnologia prin care să atingă efectiv aceste limite ultime – pentru că o poate face, în capitolul de faţă vom vedea care sunt limitele, făcându-ne astfel o idee despre cum ar putea arăta viitorul pe termen lung al vieţii. Aceste limite se bazează pe cunoştinţele noastre actuale de fizică, deci trebuie privite ca o eventuală subestimare a posibilităţilor: descoperiri ştiinţifice ulterioare ar putea sugera căi de depăşire a lor.

Ştim însă cu certitudine că viaţa viitoare va fi atât de ambiţioasă? Nu, nu ştim: poate că va deveni la fel de autosatisfacută ca un consumator de heroină sau ca un snob care nu face decât să privească la nesfârşit reluări ale episoadelor din *Keeping Up with the Kardashians*‡‡‡‡‡‡ Avem totuşi motive să bănuim că ambiţia e o trăsătură universală a vieţii avansate. Aproape indiferent ce încearcă să sporească la maximum – inteligenţă, longevitate, cunoaştere sau experienţe interesante va avea nevoie de

‡‡‡‡‡‡ Reality-show american care prezintă viaţa familiei Kardashian. *(N.* t)

resurse. Va fi deci stimulată să împingă tehnologia până la limitele ultime, pentru a obţine cât mai mult din resursele de care dispune. Apoi, singura cale de a merge mai departe este să dobândească noi resurse, extinzându-se în regiuni tot mai mari din cosmos.

Viaţa ar putea însă apărea în mod independent în mai multe locuri din univers. în acest caz, civilizaţiile neambiţioase ar deveni pur şi simplu irelevante cosmic, părţi din ce în ce mai mari din zestrea cosmică fiind în cele din urmă luate în stăpânire de forme de viaţă mai ambiţioase. Selecţia naturală se manifestă deci la scară cosmică, iar după o vreme aproape toate formele de viaţă existente vor fi ambiţioase. Pe scurt, dacă vrem să aflăm în ce măsură cosmosul nostru poate căpăta viaţă în cele din urmă, trebuie să cercetăm limitele pe care legile fizicii le impun asupra ambiţiei. Să explorăm mai întâi limitele a ceea ce putem face cu resursele (materie, energie etc.) de care dispunem în sistemul nostru solar, pentru a vedea apoi cum putem obţine mai multe resurse prin explorare şi colonizare cosmică.

FOLOSIREA LA MAXIMUM A RESURSELOR DE CARE DISPUI

Dacă supermarketurile şi bursele de mărfuri de azi vând zeci de mii de produse pe care le-am putea numi „resurse“, viaţa din viitor care a atins limitele tehnologice are în principal nevoie de o resursă fundamentală: aşa-numita *materie barionică,* adică tot ce e alcătuit din atomi sau constituenţii lor (cuarci şi electroni). Indiferent sub ce formă se găseşte această materie, tehnologia avansată o poate rearanja în orice substanţe sau obiecte se doreşte – de pildă, în centrale electrice, calculatoare sau forme superioare de viaţă. Să începem deci prin a examina limitele impuse asupra

energiei care alimentează viaţa avansată şi asupra procesării informaţiei care îi permite să gândească.

Construirea sferelor Dyson

Dacă vine vorba despre viitorul vieţii, unul dintre cei mai optimişti vizionari este Freeman Dyson. Am onoarea şi plăcerea să-l cunosc de două decenii, dar la prima noastră întâlnire am fost foarte emoţionat Eram un tânăr cercetător postdoctoral şi mâneam împreună cu prietenii mei Ia cantina Institutului pentru Studii Avansate de la Princeton, când brusc a apărut acest celebru fizician, care pe vremuri obişnuia să stea de vorbă cu Einstein şi Godel, s-a prezentat şi ne-a întrebat dacă poate să ni se alăture! M-a făcut însă imediat să mă simt în largul meu, explicându ne că preferă să ia prânzul cu tinerii decât cu bătrânii profesori pedanţi. Deşi în momentul în care scriu aceste cuvinte a ajuns la 93 de ani, Freeman are un spirit mai tânăr decât majoritatea celor pe care-i cunosc, iar lucirea poznaşă a ochilor săi arată cât de puţin ii pasă de formalităţi, ierarhii academice sau înţelepciunea convenţională. Cu cât ideea e mai îndrăzneaţă, cu atât se pasionează mai tare.

Când am discutat despre folosirea energiei, a ridiculizat lipsa de ambiţie omenească, arătând că am putea satisface toate nevoile noastre energetice actuale prin captarea luminii solare care ajunge pe o suprafaţă mai mică de 0,5% din Sahara. Dar de ce să ne oprim aici? De ce să ne oprim chiar şi la captarea întregii lumini solare care ajunge pe Pământ, lăsând să se irosească în spaţiul gol cea mai mare parte a ei? De ce să nu punem pur şi simplu toată producţia de energie a Soarelui în folosul vieţii?

Inspirat de romanul SF *Star Maker* al lui Olaf Stapledon din 1937, cu inele de lumi artificiale orbitând în jurul stelei lor, Freeman Dyson a publicat în 1960

descrierea a ceea ce avea să fie cunoscut drept *sfera Dyson.*[1] Ideea lui era de a rearanja planeta Jupiter într-o biosferă sub forma unui strat sferic înconjurând Soarele, în care descendenţii noştri s-ar dezvolta având la dispoziţie de 100 de miliarde de ori mai multă biomasă şi de un bilion de ori mai multă energie decât foloseşte omenirea în prezent.[2] Dyson susţinea că acesta e următorul pas firesc: „Este de aşteptat ca, la câteva mii de ani după ce a intrat în epoca industrială, orice specie inteligentă să ocupe o biosferă artificială care înconjoară complet steaua ei.“ Dacă locuieşti în interiorul unei sfere Dyson, nu există nopţi, vezi venind de pretutindeni lumina Soarelui reflectată de restul biosferei, aşa cum vezi azi lumina Soarelui reflectată de Lună. Dacă vrei să vezi stelele, trebuie să „urci“ şi să priveşti cosmosul din afara sferei Dyson.

O soluţie presupunând tehnologie de nivel mai modest pentru a construi o sferă Dyson parţială este plasarea unui inel de habitate pe o orbită circulară în jurul Soarelui. Pentru a înconjura complet Soarele, adaugi inele sub axe diferite şi la distanţe uşor diferite, ca să eviţi ciocnirile. Dificultatea de a conecta între ele aceste inele aflate în mişcare rapidă, ceea ce complică transportul şi comunicarea, poate fi depăşită dacă se construieşte o sferă Dyson staţionară monolitică, la care atracţia gravitaţională a Soarelui e echilibrată de presiunea radiaţiei solare – idee propusă de Robert L Forward şi Colin Mclnnes. Sfera poate fi construită adăugând treptat „sateliţi“: sateliţi staţionari care contrabalansează gravitaţia Soarelui prin presiunea radiaţiei, nu prin forţe centrifuge. Ambele forţe scad cu pătratul distanţei până la Soare, ceea ce înseamnă că dacă pot fi echilibrate Ia o distanţă faţă de Soare, ele se vor echilibra la orice distanţă, sateliţii putând fi plasaţi oriunde în sistemul solar. Sateliţii trebuie să fie nişte foi extrem de uşoare, cântărind doar 0,77 grame de metru

pătrat, de vreo sută de ori mai puţin decât hârtia, dar asta nu constituie un impediment. De exemplu, o foaie de grafen (un singur strat de atomi de carbon aranjaţi într-o structură hexagonală de tip fagure) cântăreşte de o mie de ori mai puţin decât acea greutate-limită. Dacă sfera Dyson e construită aşa încât să nu absoarbă, ci să reflecte cea mai mare parte a luminii solare, atunci creşte intensitatea reflectării luminii, sporind presiunea radiaţiei şi făcând astfel posibil ca sfera să devină şi mai grea. Multe stele au o luminozitate de o sută, ba chiar de o mie de ori mai mare decât Soarele nostru, deci pot suporta sfere Dyson cu atât mai grele.

Dacă vrem să avem în sistemul nostru solar o sferă Dyson rigidă mult mai grea, atunci pentru a se opune gravitaţiei Soarelui e nevoie de materiale ultrarezistente care pot suporta presiuni de zeci de mii de ori mai mari decât acelea din structura de rezistenţă a celor mai înalţi zgârie-nori, care nu se topesc şi nu se deformează. Pentru a dăinui multă vreme, o sferă Dyson trebuie să fie dinamică şi inteligentă, reglându-şi constant poziţia şi forma ca răspuns la perturbaţii, şi deschizând găuri largi pentru a lăsa asteroizii şi cometele să treacă prin ea fără incidente. Se poate la fel de bine ca un sistem de detecţie şi deviere să fie folosit pentru face faţă acestor intruşi, eventual dezmembrându-i şi utilizându-le mate ria în mod convenabil.

Pentru oamenii de azi, viaţa pe o sferă Dyson sau în interiorul ei ar fi în cel mai bun caz derutantă, iar în cel mai rău caz imposibilă, dar asta nu va împiedica viitoarele forme de viaţă, biologice sau nebiologice, să se dezvolte acolo. în varianta cu sateliţi, practic n-ar exista gravitaţie, iar dacă te-ai plimba pe un satelit staţionar ai merge doar pe suprafaţa orientată spre exterior (în direcţia opusă Soarelui) fără să cazi, sub acţiunea unei gravitaţii de vreo zece mii de ori mai slabă decât cea cu care eşti obişnuit. N-ar exista câmp magnetic (cu excepţia cazului în care tu

îl creezi) şi ai fi protejat de particulele periculoase emise de Soare. Avantajul ar fi că o sferă Dyson de dimensiunea orbitei actuale a Pământului ne-ar pune la dispoziţie o suprafaţă de vreo 500 de milioane de ori mai mare decât cea pe care trăim azi.

Dacă ne dorim habitate cu condiţii asemănătoare celor de pe Pământ, vestea bună este că ele sunt mult mai uşor de construit decât o sferă Dyson. De exemplu, figurile 6.1 şi 6.2 prezintă planul unui habitat cilindric conceput de fizicianul american Gerard K. O'Neil, cu gravitaţie artificială, protecţie contra radiaţiei cosmice, ciclu zi-noapte de 24 de ore şi atmosferă şi ecosisteme de tipul celor de pe Pământ. Asemenea habitate ar putea orbita liber în interiorul unei sfere Dyson, sau variante modificate ar putea fi ataşate în exteriorul ei.

Construirea unor centrale electrice mai eficiente

Deşi sferele Dyson sunt eficiente energetic în raport cu standardele de azi, ele împing doar cu puţin limitele impuse de legile fizicii. Einstein ne-a arătat că dacă vrem să convertim masa în energie cu o eficienţă de 100%,[§§§§§§§] atunci o cantitate de masă m ne-ar

[§§§§§§§] Dacă lucrezi în sectorul energetic eşti pesemne obişnuit să defineşti eficienţa drept fracţiunea din energia eliberată care se află într-o formă utilă. *(N.a.)*

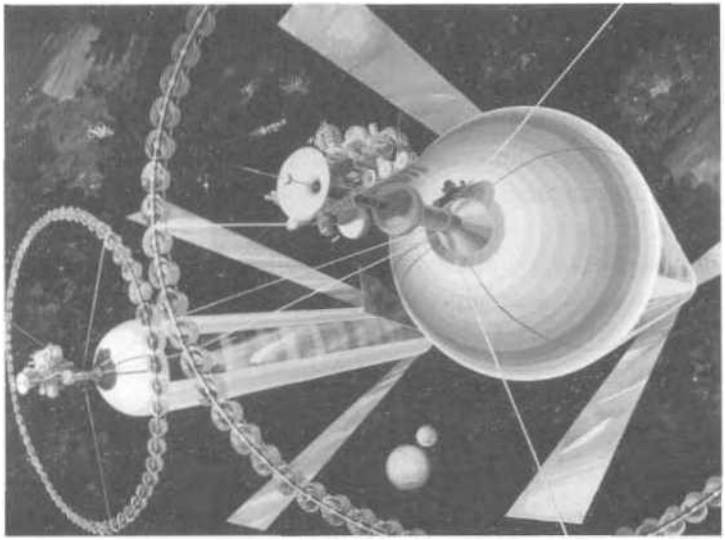


Figura 6.10 pereche de cilindri O'Neilîn contrarotaţie ar oferi habitate confortabile de tip terestru dacă ar orbita în jurul Soarelui astfel încât să fie orientate mereu drept spre el. Forţa centrifugă a rotaţiei lor produce o gravitaţie artificială, iar trei oglinzi pliabile proiectează în interior lumina solară într-un ciclu zi-noapte de 24 de ore. Habitatele mici di spuse într-un inel sunt destinate agriculturii. Imagine obţinută prin amabilitatea lui RickGuidice/NASA.

da o cantitate de energie *E*conform celebrei formule $E=mc^2$, unde *c* este viteza luminii. Asta înseamnă că, *c* fiind foarte mare, o cantitate mică de masă poate produce o cantitate uriaşă de energie. Dacă am avea rezerve importante de antimaterie (dar n-avem), atunci ar fi uşor de construit o centrală electrică 100% eficientă: turnând o linguriţă de anti-apă în apă obişnuită s-ar elibera energia echivalentă cu cea a 200 000 de tone de TNT, corespunzând unei bombe cu hidrogen tipice – suficientă pentru a satisface nevoile de energie ale întregii lumi pentru circa şapte minute.

în prezent însă, cele mai răspândite modalităţi de generare a energiei sunt jalnic de ineficiente, după cum se vede în tabelul 6.r şi în figura 6.3. Digerarea unei bomboane are o eficienţă de doar 0,00000001%, în

sensul că eliberează doar o zecime de bili- onime din energia mc^2 pe care o conţine. Dacă stomacul tău ar

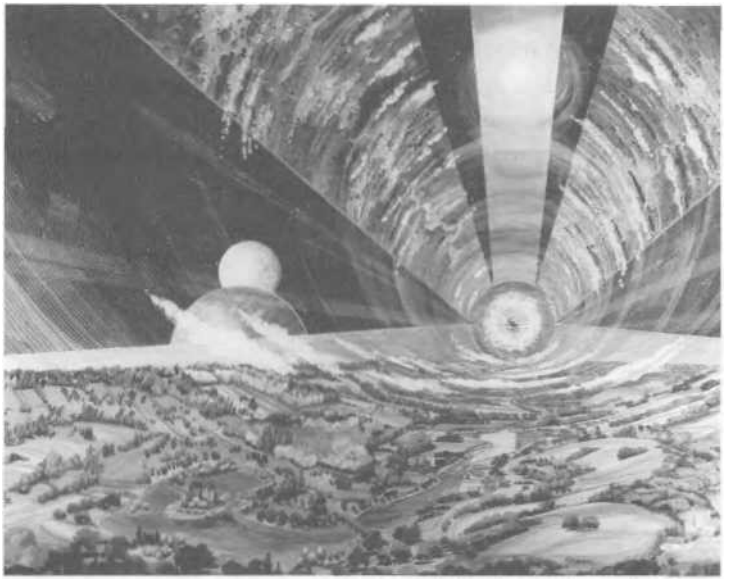


Figura *6.2* Interiorul unuia dintre cilindrii O'Neil din figura precedentă. Dacă diametrul lui este de 6,4 km şl efectuează o rotaţie la fiecare 2 minute, oamenii vor simţi aceeaşi gravitaţie aparentă ca pe Pământ. Soarele se află în spate, dar pare să fie deasupra graţie unei oglinzi din exteriorul cilindrului care se pliază noaptea. Ferestre ermetice împiedică atmosfera să iasă din cilindru. Imagine obţinută prin amabilitatea lui Rick Guidice/NASA.

avea o eficienţă chiar şi de 0,001%, ai avea nevoie să mănânci o singură dată pentru tot restul vieţii. în comparaţie cu mâncatul, arderea cărbunelui şi benzinei este de doar 3, respectiv 5 ori mai eficientă. Reactoarele nucleare de azi se descurcă mult mai bine despicând atomii de uraniu prin fisiune, dar tot nu reuşesc să extragă mai mult de 0,08% din energia lor. Reactorul nuclear din miezul Soarelui este cu un ordin de mărime mai eficient, extră- gând 0,7% din fuziunea hidrogenului pentru a produce heliu. Chiar dacă am înconjura Soarele cu o sferă perfectă, am converti abia 0,08% din masa Soarelui în energie pe care s-o putem folosi, fiindcă atunci când Soarele îşi va consuma circa o zecime din hidrogenul său, îşi va încheia viaţa de stea normală, devenind o

gigantă roşie şi începând să moară. Lucrurile nu stau mai bine nici pentru alte stele: fracţiunea de hidrogen consumată în decursul vieţii lor principale variază între 4%, la stelele foarte mici, şi 12%, la stelele cele mai mari. Dacă am construi un reactor de fuziune care să folosească 100% din hidrogenul pe care îl avem la dispoziţie, am rămâne tot la eficienţa de 0,7% a procesului de fuziune. Cum ne-am putea descurca mai bine?

Metodă	**Eficienţă**
Digerarea unei bomboane	0,00000001%
Arderea cărbunelui	0,00000003%
Arderea benzinei	0,00000005%
Fisiunea uraniului-235	0,08%
Folosirea unei sfere Dyson până când Soarele moare	0,08%
Fuziunea hidrogenului pentru a produce heliu	0,7
Motor cu gaură neagră în rotaţie	29%
Sferă Dyson în jurul unui quasar	42%
Sfa Ieri zor	50%?
Evaporarea găurilor negre	90%

Tabelul 6.1 Eficienţa convertirii masei în energie utilizabilă, în raport cu limita teoretică $E = mc^2$. După cum se arată în text, procesul în care hrăneşti găurile negre şi aştepţi ca ele să se evapore, cu o eficienţă energetică de 90%, este din păcate prea lent pentru a fi util, iar accelerarea procesului îi scade mult eficienţa.

Evaporarea găurilor negre

în cartea sa *O scurtă istorie a timpului,* Stephen Hawking propune ideea de centrală electrică pe bază de găuri negre.******** Poate părea paradoxal, odată ce multă vreme s-a crezut că găurile negre sunt capcane din care nimic nu poate vreodată evada, nici măcar lumina. Hawking

******** Dacă în universul apropiat nu se află nici o gaură neagră naturală potrivită, poţi crea una inghesuind o cantitate mare de materie într-un spaţiu suficient de mic. *(N. a.)*

însă a calculat că efectele gravitaţiei cuantice fac ca o gaură neagră să se comporte ca un obiect fierbinte – cu cât

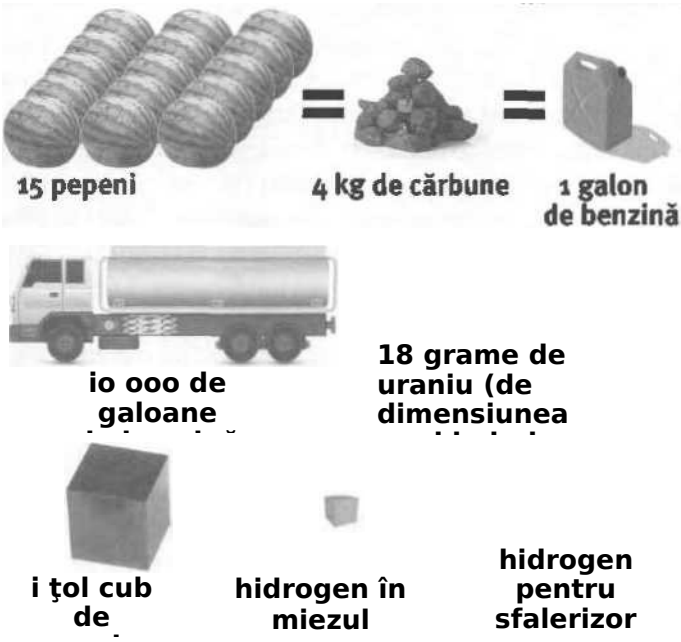


Figura 6.3 Tehnologia avansată poate extrage mult mai multă energie din materie decât obţinem mâncând sau arzând-o, iar până şi fuziunea extrage de 140 de ori mai puţină energie decât am extrage dacă ne-am situa la limitele impuse de fizică. Centrale eletrice exploatând sfaleroni, quasari sau găuri negre care se evaporă ar fi mult mai eficiente.

mai mic, cu atât mai fierbinte – care emite o radiaţie termică numită acum *radiaţie Hawking.* Asta înseamnă că, treptat, gaura neagră pierde energie şi se evaporă. Cu alte cuvinte, indiferent ce materie arunci în gaura neagră, ea se va întoarce sub formă de radiaţie termică, aşa încât, atunci când gaura neagră s-a evaporat complet, ai convertit materia în radiaţie cu o eficienţă de aproape 100%.††††††††

†††††††† Este o supraestimate, fiindcă radiaţia Hawking conţine şi particule de la care e greu de extras lucru mecanic util. Găurile negre mari au o eficienţă de doar 90%, pentru că aproximativ 10% din energie e radiată sub formă de gravitoni: particule extrem de „timide“, care sunt aproape imposibil de detectat, ca să nu mai vorbim de extragerea lucrului mecanic util. Pe măsură ce gaura neagră continuă să se evapore şi să se mişcoreze, eficienţa scade şi

Dacă nu cumva gaura neagră are dimensiuni mai mici decât un atom, este problematică folosirea evaporării ei ca sursă de energie, pentru că procesul e neverosimil de lent – durează mai mult decât vârsta actuală a universului iar gaura neagră radiază mai puţină energie decât o lumânare. Puterea produsă scade cu pătratul dimensiunii găurii negre, aşa încât fizicienii Louis Crane şi Shawn Westmoreland au propus folosirea unei găuri negre de vreo mie de ori mai mică decât un proton, cântărind cam cât cel mai greu vapor.[3] Ideea lor era ca motorul cu gaură neagră să fie folosit la propulsia unei nave spaţiale (subiect asupra căruia vom reveni), aşa încât erau preocupaţi mai curând de portabilitate decât de eficienţă, şi au propus ca gaura neagră să fie alimentată cu lumină laser, fără să existe vreo conversie energie-materie. Chiar dacă ai alimenta-o cu materie în loc de radiaţie, este dificil să garantezi obţinerea unei eficienţe înalte: pentru a intra într-o gaură neagră de o mie de ori mai mică decât ei, protonii trebuie proiectaţi asupra găurii negre cu o instalaţie la fel de puternică pe cât este Marele Accelerator de Hadroni‡‡‡‡‡‡‡, crescând energia lor mc^1 cu cel puţin de o mie de ori mai multă energie cinetică (de mişcare). Cum cel puţin 10% din acea energie cinetică s-ar pierde prin gravitoni atunci când gaura neagră se evaporă, am introduce mai multă energie în gaura neagră decât am extrage, obţinând o eficienţă negativă. Altă piedică în calea construirii unei centrale cu găuri negre ţine de faptul că nu avem încă o teorie riguroasă a gravitaţiei cuantice pe care să ne bazăm calculele – dar această incertitudine Iasă deschisă posibilitatea să existe efecte utile ale gravitaţiei cuantice pe care deocamdată nu le-am descoperit.

---

mai mult, fiindcă radiaţia Hawking începe să includă neutrini şi particule masive. (N. *a.)*

‡‡‡‡‡‡‡ Large HadronCollider(LHC) -acelerator de particule construit Ia Centrul European pentru Cercetări Nucleare (CERN), în apropiere de Geneva. *(N. L)*

Există din fericire alte căi de a folosi găurile negre în centrale, fără a implica gravitaţia cuantică sau alte aspecte puţin înţelese ale fizicii. De pildă, multe găuri negre se învârt foarte repede în jurul axei proprii, orizontul evenimentelor rotindu-se cu o viteză apropiată de cea a luminii, iar această energie de rotaţie poate fi extrasă. Orizontul evenimentelor unei găuri negre este regiunea din care nici măcar lumina nu poate evada, pentru că atracţia gravitaţională e prea puternică. Figura 6.4 arată că, pentru o gaură neagră în rotaţie, în afara orizontului evenimentelor există o regiune numită *ergosferă,* în care gaura neagră antrenează în mişcare spaţiul atât de repede, încât nici o particulă nu poate rămâne în repaus, fiind şi ea antrenată în mişcare. Dacă arunci un obiect în

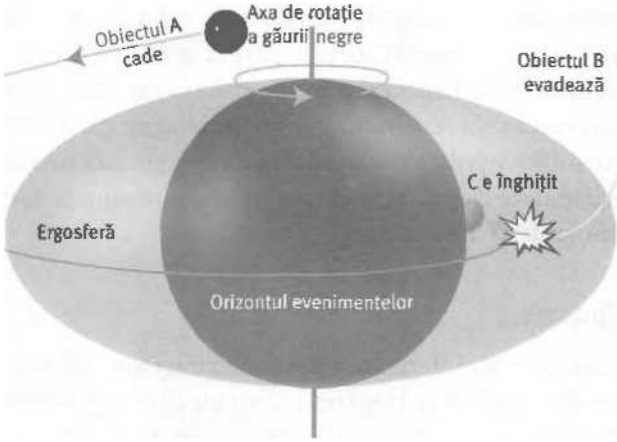


Figura 6.4 O parte din energia de rotaţie a unei găuri negre care seînvârteîn jurul axei proprii poate fi extrasă aruncând o particulă Aîn apropierea găurii negre şi făcând-o să se rupă într-o parte C care e înghiţită şi o parte B care evadează – cu energie mai mare decât a avut

ergosferă, acesta va fi accelerat într-o mişcare rapidă în jurul găurii negre. Din păcate, va fi curând înghiţit de ea, dispărând pentru totdeauna prin orizontul evenimentelor, aşadar nu poţi extrage astfel energie. Totuşi, Roger Penrose a descoperit că dacă lansezi obiectul sub un anumit unghi şi faci aşa încât să se rupă în două bucăţi, după cum se vede în figura 6.4, se poate ca doar o bucată să fie înghiţită, iar cealaltă să se îndepărteze de gaura neagră cu o energie mai mare decât cea cu care ai lansat obiectul iniţial. Altfel spus, ai convertit o parte din energia de rotaţie a găurii negre în energie pe care o poţi folosi. Repetând de multe ori procesul, poţi extrage *toată* energia de rotaţie a găurii negre, aşa încât aceasta încetează să se rotească, iar ergosfera ei dispare. Dacă gaura neagră iniţială se roteşte pe cât de repede permit legile naturii, orizontul evenimentelor deplasându-se practic cu viteza luminii, prin această strategie ai putea converti în energie 29% din masa ei. Nu se ştie încă precis cât de repede se rotesc găurile negre observate, dar multe dintre cele atent studiate par să se rotească foarte repede: între 30% şi 100% din maximumul permis. Gaura neagră uriaşă din centrul galaxiei noastre (de 4 milioane de ori mai grea decât Soarele) pare să se rotească, deci chiar dacă numai 10% din masa ei ar fi convertită în energie utilă, s-ar obţine cam ceea ce ar rezulta din convertirea în energie a 400 000 de sori cu o eficienţă de 100%, sau tot atâta energie câtă am extrage de la nişte sfere Dyson înconjurând 500 de milioane de sori, timp de miliarde de ani.

Quasari

Altă strategie interesantă este de a extrage energie nu de la gaura neagră însăşi, ci de la materia care cade în ea. Natura a găsit deja o modalitate de a face asta – prin quasari. Pe măsură ce un gaz se roteşte în apropierea unei găuri negre,

formând un disc ale cărui părţi centrale sunt treptat înghiţite, gazul devine extrem de fierbinte şi emite mari cantităţi de radiaţie. Când gazul cade spre gaura neagră, e accelerat, transformând energia sa potenţială gravitaţională în energie cinetică. Turbulenţe complexe fac ca mişcarea coordonată a masei de gaz să devină, la scară mică, tot mai dezordonată, până când atomii individuali încep să se ciocnească între ei la viteze mari – agitaţia aceasta înseamnă temperatură ridicată, iar ciocnirile violente convertesc energia de mişcare în radiaţie. Dacă am construi o sferă Dyson în jurul întregii găuri negre, la o distanţă sigură, această energie a radiaţiei ar putea fi captată şi folosită. Cu cât se roteşte mai rapid gaura neagră, cu atât procesul devine mai eficient, la viteza de rotaţie maximă a găurii negre eficienţa ajungând la 42%.[§§§§§§§§] Pentru găurile negre de greutatea unei stele, majoritatea energiei e eliberată sub formă de raze X, iar pentru găurile negre supermasive, de genul celor din centrul galaxiilor, o mare parte a energiei se află în domeniile luminii infraroşii, vizibile şi ultraviolete.

Când nu mai ai cu ce alimenta gaura neagră, poţi începe să extragi energia ei de rotaţie, aşa cum am văzut mai sus.[********] Şi în acest caz natura a găsit deja o soluţie,

---

§§§§§§§§ Fanii lui Douglas Adams ar trebui să observe că aceasta e o întrebare elegantă al cărei răspuns coincide cu cel la întrebarea privind viaţa, universul şi tot ce există. Mai precis, eficienţa este 1 – V3 a 42%. (N. *a.)* [în seria de romane a lui Douglas Adams *Ghidul autostopistului galactic,* răspunsul ultim dat de un supercalculator este pur şi simplu numărul 42, dar supercalculatonil nu mai poate spune apoi şi care e întrebarea ultimă – n.t]

********♦* Dacă alimentezi gaura neagră plasând în jurul ei un nor de gaz care se roteşte lent în acelaşi sens cu ea, atunci gazul se va învârti tot mai repede şi va fi înghiţit de ea, accelerând rotaţia găurii negre, la fel cum un patinator se învârte mai repede când îşi apropie braţele de corp. în felul acesta rotaţia găurii negre poate fi menţinută la viteză maximă, permiţându-ţi să extragi întâi

intensificând radiaţia unei mase de gaz printr-un proces magnetic numit mecanism Blandford-Znajek. Nu-i exclus ca tehnologia să ne ajute să obţinem o eficienţă în extragerea energiei de peste 42%, folosind în mod inteligent câmpuri magnetice şi alte elemente.

Sfaleroni

Există şi o modalitate de a converti materia în energie fără a implica găurile negre, printr-un proces implicând *şfaleronii.* Acesta poate distruge cuarcii, transformându-i în leptoni: electroni, miuoni şi particule tau (verişorii mai grei al electronilor), neutrini, precum şi antiparticulele corespunzătoare tuturor acestora.[4] După cum se arată în figura 6.5, modelul standard al fizicii particulelor prezice că nouă cuarci cu anumite valori ale aromei şi spinului†††††††††† se pot uni şi transforma în trei leptoni, trecând printr-o stare intermediară numită sfaleron. Pentru că masa totală a ceea intră în reacţie este mai mare decât masa totală a ceea ce rezultă, diferenţa de masă este convertită în energie conform formulei lui Einstein $E = mc^1$.

42% din energia gazului, iar apoi 19% din ce rămâne, eficienţa totală fiind 42°to + (1 – 42%) x 29% = 59%. *(N. a.)*

†††††††††† Spinul este o caracteristică a particulelor, asociată la începuturile mecanicii cuantice cu rotaţia particulei în jurul axei proprii (de la englezescul *to spin-*„a se învârti"). Aroma *(flavor)* este o caracteristică ce defineşte mai multe tipuri de cuarci şi leptoni. *(N.* t)

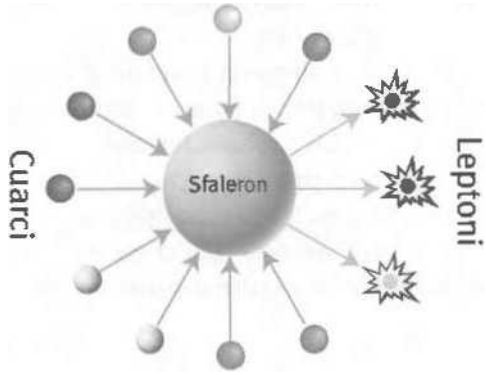


Figura 6.5 Conform modelului standard al fizicii particulelor, nouă cuarci cu anumite valori ale aromei şi spinului se pot uni şi transforma în trei leptoni, trecând printr-o stare intermediară numită sfaleron. Suma maselor cuarcilor (împreună cu energia gluonilor‡‡‡‡‡‡‡‡‡ ca re-¡însoţesc) e mult mai mare decât suma maselor leptonilor, prin urmare în acest proces se va elibera energie, ilustrată grafic prin scântei.

Viaţa inteligentă din viitor ar putea deci construi ceea ce eu numesc un *sfalerizor*. un generator de energie acţionând ca un motor diesel hipertrofiat. Motorul diesel tradiţional comprimă un amestec de aer şi motorină până când temperatura creşte suficient pentru ca arderea să se declanşeze spontan, după care amestecul fierbinte se destinde şi efectuează lucru mecanic util – de pildă, împingând un piston. Dioxidul de carbon şi alte gaze de combustie cântăresc cu aproximativ 0,00000005% mai puţin decât amestecul aflat iniţial în cilindru, iar diferenţa de masă se transformă în energie termică ce acţionează asupra pistonului. Un sfalerizor ar comprima materia obişnuită până la câteva trilioane (10^{15}) de grade, pentru a o lăsa apoi să se destindă şi să se răcească, odată ce sfaleronii şi-au îndeplinit misiunea.§§§§§§§§ Cunoaştem deja rezultatul acestui

‡‡‡‡‡‡‡‡‡ Gluonii (de la englezescul glue – „lipici“) sunt particulele care mijlocesc interacţiunile nucleare tari dintre cuarci, constituenţi fundamentali ai materiei. (IV. t)

§§§§§§§§§ Materia trebuie să fie suficient de fierbinte pentru a

experiment, fiindcă universul a trecut prin aşa ceva în urmă cu vreo 13,8 miliarde de ani, pe când era extrem de fierbinte: aproape 100% din materie s-a transformat în energie, rămânând mai puţin de o miliardime din particule pentru a alcătui materia obişnuită compusă din cuarci şi electroni. Sfalerizorul seamănă deci cu un motor diesel, numai că e de peste un miliard de ori mai eficient! Alt avantaj este că nu contează ce combustibil introduci – funcţionează cu orice e compus din cuarci, deci cu orice materie obişnuită.

Datorită acestor procese Ia temperatură înaltă, în universul nostru timpuriu a apărut de peste un bilion de ori mai multă radiaţie (fotoni şi neutrini) decât materie (cuarci şi electroni, care mai târziu s-au agregat în atomi). în următorii 13,8 miliarde de ani s-a produs o netă separare, atomii s-au concentrat în galaxii, stele şi planete, iar majoritatea fotonilor au rămas în spaţiul inter- galactic, formând radiaţia cosmică de fond de microunde, prin intermediul căreia am căpătat o imagine a universului timpuriu. Orice formă avansată de viaţă care trăieşte într-o galaxie sau în altă concentrare de materie poate deci transforma cea mai mare parte din materia disponibilă înapoi în energie, readucând procentajul materiei la aceeaşi valoare minusculă de la începutul universului prin re crearea pentru scurt timp a acelor condiţii în interiorul unui sfalerizor.

Ca să ne dăm seama cât de eficient ar fi un sfalerizor, avem nevoie să cunoaştem unele detalii practice esenţiale: de pildă, cât de mare trebuie să fie pentru a împiedica scurgerea unei fracţiuni importante din fotoni şi neutrini în faza de compresie? Ceea ce putem totuşi spune cu certitudine este că în viitor viaţa va beneficia de posibilităţi

---

reunifica forţele electromagnetică şi nucleară slabă, ceea ce se întâmplă atunci când particulele se deplasează cu viteze corespunzând, intr-un accelerator de particule, aplicării unei tensiuni de zoo de miliarde de volţi. *(N. a.)*

energetice incomparabil superioare celor permise de tehnologia actuală. N-am reuşit nici măcar să construim un reactor de fuziune, dar tehnologia viitorului ar trebui să poată fi de zece, ba chiar de o sută de ori mai eficientă.

Calculatoare mai performante

Dacă eficienţa energetică a digestiei este de io miliarde de ori mai mică decât limita fizică a convertirii masei în energie, cât de eficiente sunt calculatoarele de azi? După cum vom vedea, lucrurile stau încă şi mai rău.

Spun adesea despre prietenul meu Seth Lloyd că e pesemne singurul coleg de la MIT care e la fel de nebun ca mine. După cercetări de pionierat în domeniul calculatoarelor cuantice, a scris o carte în care susţinea că întregul univers e un calculator cuantic. După muncă, obişnuim să mergem la o bere, şi n-am găsit încă un subiect despre care să nu aibă de spus ceva interesant. De pildă, aşa cum am menţionat în capitolul 2, are multe de spus despre limitele ultime ale calculului. într-un celebru articol publicat în anul 2000, a demonstrat că viteza de calcul e limitată de energie: efectuarea unei operaţii logice elementare în timpul *T*necesită o energie medie $E = hJ\hat{}T$, unde *h* este mărimea fundamentală din fizică numită constanta lui Planck. Asta înseamnă că 1 kilogram de calculator poate efectua cel mult 5 x 10^{50} operaţii pe secundă – cu 36 de ordine de mărime mai mult decât laptopul la care scriu aceste cuvinte! Vom atinge această limită în câteva secole, dacă puterea de calcul va continua să se dubleze la fiecare doi ani, după cum am văzut în capitolul 2. El a arătat de asemenea că 1 kilogram de calculator poate stoca cel mult 10^{31} biţi, adică de circa un miliard de miliarde de ori mai mult decât laptopul meu.

Seth a fost primul care a înţeles că atingerea acestor limite ar fi dificilă chiar şi pentru viaţa superinteligentă, fiindcă memoria acelui i kilogram de „calculator“ ultim ar semăna cu o explozie termonucleară sau cu un mic fragment din big bang. El crede însă că limitele practice nu sunt foarte departe de cele ultime. Prototipurile de calculatoare cuantice existente şi-au miniaturizat deja memoria stocând un bit per atom, ceea ce la scară mare ar permite stocarea a circa ro^{25} biţi/kg – de un bilion de ori mai mult decât laptopul meu. Apoi, folosind radiaţia electromagnetică pentru comunicarea între aceşti atomi, s-ar ajunge la aproximativ 5 x io^{40} operaţii pe secundă – cu 31 de ordine de mărime mai mult decât unitatea centrală de procesare a laptopului meu.

Pe scurt, potenţialul de calcul al vieţii din viitor este intr-adevăr uluitor: în termenii ordinelor de mărime, cele mai bune super- computere de azi se află mult mai departe de acel 1 kg de calculator ultim decât de semnalizatorul unei maşini, un dispozitiv care stochează doar un bit de informaţie, oscilând între conectat şi deconectat o dată pe secundă.

Alte resurse

Din perspectiva fizicii, orice ar vrea să creeze viaţa viitoare – de la habitate şi maşini la noi forme de viaţă – înseamnă pur şi simplu aranjarea particulelor elementare într-un anumit mod. La fel cum o balenă albastră este mici crustacee rearanjate, iar micile crustacee sunt plancton rearanjat, întregul nostru sistem solar nu-i decât hidrogen rearanjat în cursul a 13,8 miliarde de ani de evoluţie cosmică: gravitaţia a rearanjat hidrogenul în stele, care au rearanjat hidrogenul în atomi mai grei, după care gravitaţia a rearanjat aceşti atomi în planeta noastră, unde procesele chimice şi biologice i-au rearanjat în viaţă.

Viaţa viitorului care şi-a atins limitele tehnologice poate efectua asemenea rearanjări ale particulelor mult

mai rapid şi eficient, folosindu-şi în primă instanţă puterea de calcul pentru a găsi metoda optimă, iar apoi energia disponibilă pentru a alimenta procesul de rearanjare a materiei. Am văzut că materia poate fi

convertită atât în calculatoare, cât şi în energie, aşa încât, într-un anumit sens, este singura resursă fundamentală de care e nevoie.********** Odată ce viaţa viitorului a atins limitele fizice pentru ceea ce poate face cu materia, îi rămâne o singură soluţie de a face mai mult: să obţină mai multă materie. Iar pentru asta singura soluţie este să se extindă în univers. Cu toată viteza spre spaţiu!

OBŢINEREA RESURSELOR PRIN COLONIZAREA COSMOSULUI

Cât de mare e zestrea noastră cosmică? Mai precis, ce limite superioare impune fizica asupra cantităţii de materie pe care viaţa ar putea-o în cele din urmă folosi? Desigur, zestrea noastră cosmică e uluitor de mare, dar, în mod exact, cât de mare' Tabelul 6.2 prezintă câteva cifre semnificative. Planeta noastră este acum 99,999999% moartă, în sensul că acest procentaj din materia ei nu face parte din biosferă şi nu face aproape nimic util pentru viaţă, în afara faptului *că oferă* atracţie gravitaţională şi un câmp magnetic. Există deci în principiu posibilitatea ca într-o bună zi să fie folosită de o sută de milioane de ori mai multă materie în beneficiul vieţii. Dacă am putea utiliza optim materia din sistemul nostru solar (inclusiv Soarele), ne-am descurca de încă o sută de milioane de ori mai bine. Colonizarea galaxiei ne-ar spori resursele de încă un bilion de ori.

********** Am vorbit doar despre materia alcătuită din atomi. Există de vreo şase ori mai multă materie întunecată, dar e foarte greu de detectat şi captat, aşa încât rămâne de văzut dacă viaţa viitoare o va putea utiliza. *(N.* o.) [Tot ce ştim despre materia întunecată este că, din motive de ordin astrofizic, ar trebui să existe, dar deocamdată n-a fost observată direct, şi nici nu se cunoaşte compoziţia ei; a nu se confunda cu *energia întunecată* (responsabilă pentru expansiunea accelerată a universului) despre care va fi vorba mai jos – *n,* t]

Vă închipuiţi pesemne că putem obţine resurse nelimitate colonizând oricâte galaxii vrem, dacă suntem suficient de perseverenţi,

Regiune	Particule
Biosfera noastră	**10*3**
Planeta noastră	**10^{5}¹**
Sistemul nostru solar	**1057**
Galaxia noastră	**IO⁶’**
Zona accesibilă călătorind cu jumătate din viteza luminii	**toZS**
Zona accesibilă călătorind cu viteza luminii	**IO?⁶**
Universul nostru	**IO?⁸**

Tabelul 6.2 Numărul aproximativ al particulelor de materie (protoni şi neutroni) pe care viaţa viitoare ar putea spera să le folosească.

însă nu asta ne sugerează cosmologia modernă. Da, spaţiul însuşi s ar putea să fie infinit, conţinând infinit de multe galaxii, stele şi planete – şi într-adevăr asta prezic cele mai simple versiuni ale *inflaţiei,* paradigma ştiinţifică actuală care explică big bang-ul de acum 13,8 miliarde de ani. Dar chiar dacă există infinit de multe galaxii, vedem doar vreo 200 de miliarde şi putem coloniza cel mult 10 miliarde de galaxii.

Ceea ce ne limitează este viteza luminii: un an-lumină (aproxi mativ 10 bilioane de kilometri) pe an. Figura 6.6 prezintă partea din spaţiu de unde lumina a ajuns la noi în cursul celor 13,8 miliarde de ani scurşi de la big bang, o regiune sferică numită „universul nostru observabil” sau, mai simplu, „universul *nostru“.* Chiar dacă spaţiul e infinit, universul nostru e finit, conţinând „doar” aproxi-mativ 10^{78} atomi. Apoi, circa 98% din universul nostru „poate fi văzut, dar nu şi atins“, în sensul că-l putem vedea, dar niciodată nu vom ajunge la el, chiar dacă am călători la nesfârşit cu viteza luminii. De ce oare? La

urma urmei, limita până la care putem vedea e dată de faptul că universul nostru nu are o vechime infinită, aşa încât lumina de la mare distanţă n-a avut încă timp să ajungă la noi. N-am putea aşadar călători până la galaxii oricât de îndepărtate dacă nu ne impunem vreo limită pentru durata călătoriei?

Primul impediment ţine de faptul că universul nostru se extinde, ceea ce înseamnă că aproape toate galaxiile se îndepărtează

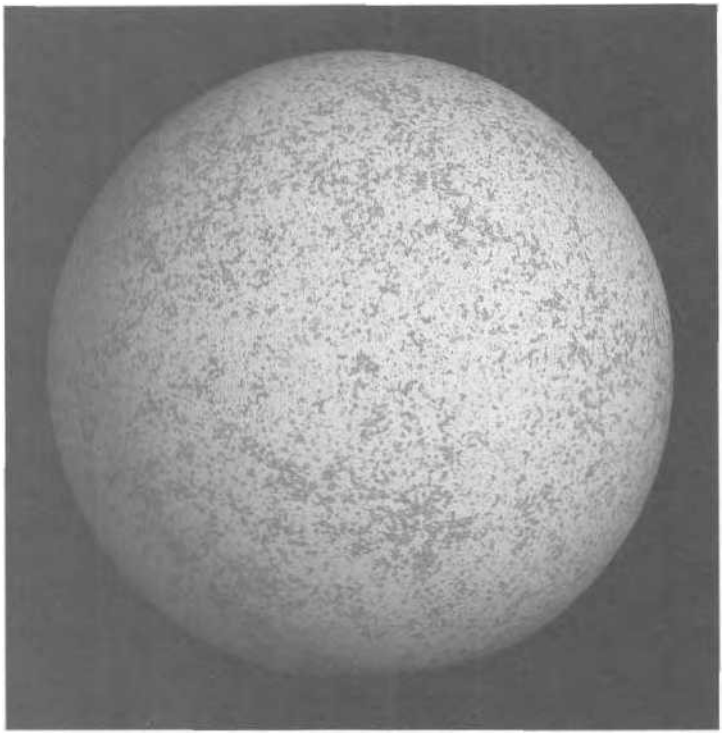


Figura 6.6 Universul nostru – regiunea sferică din spaţiu de unde lumina a avut timp să ajungă la noi (ne aflăm în centru) în cursul celor 13,8 miliarde de ani scurşi de la big bang. Structurile de pe suprafaţă sunt „fotografii de bebeluş“ ale universului obţinute de satelitul Planck şi arată că, pe când avea vârsta de doar 400 000 de ani, era alcătuit din plasmă fierbinte, cu o temperatură aproape la fel de ridicată ca cea de la suprafaţa Soarelui. Probabil că spaţiul se continuă dincolo de această regiune, iarîn fiecare an ni se dezvăluie materie nouă.

de noi, aşa încât colonizarea galaxiilor aflate la mare distanţă e un fel de joc de-a prinselea. Al doilea impediment ţine de faptul că această expansiune cosmică e accelerată, datorită misterioasei energii întunecate care contribuie cu circa 70% la energia totală a universului nostru. Pentru a înţelege ce problemă creează asta, imaginează-ţi că ajungi pe peronul gării şi vezi că trenul

tău accelerează încet îndepărtându-se de tine, dar o uşă a rămas deschisă. Dacă eşti rapid şi temerar, poţi oare prinde trenul? în cele din urmă va merge mai repede decât poţi tu alerga, deci răspunsul depinde de distanţa la care te afli iniţial faţă de tren: dacă e mai mare decât o anumită valoare critică, n-ai să poţi prinde trenul niciodată. Suntem în aceeaşi situaţie încercând să ajungem din urmă acele galaxii de la mare distanţă care se îndepărtează accelerat: chiar dacă am călători cu viteza luminii, toate galaxiile aflate la o distanţă mai mare de 17 miliarde de ani-lumină rămân pe veci de neatins – iar ele reprezintă peste 98% din galaxiile universului nostru.

Dar staţi o clipă: nu spune oare teoria relativităţii speciale a lui Einstein că nimic nu poate călători mai rapid decât lumina? Atunci cum de întrec galaxiile ceva ce călătoreşte cu viteza luminii? Răspunsul este că teoria relativităţii speciale este înlocuită de teoria relativităţii generale a lui Einstein, care e mai permisiva în privinţa limitării vitezei: nimic nu poate călători mai repede decât lumina *prin spaţiu,* dar spaţiul e liber să extindă cât de repede vrea. Einstein ne-a oferit şi o modalitate de a vizualiza aceste limite ale vitezei, considerând timpul ca a patra dimensiune a spaţiului-timp (vezi figura 6.7, în care am păstrat doar trei dimensiuni, omiţând una dintre cele trei dimensiuni spaţiale). Dacă spaţiul nu s-ar extinde, razele de lumină ar fi linii drepte înclinate la 45^0 prin spaţiu-timp, aşa încât regiunile pe care le-am putea vedea şi la care am putea ajunge de aici ar fi conuri. Conul nostru luminos din trecut ar fi trunchiat de big bang-ul din urmă cu 13,8 miliarde de ani, iar conul nostru luminos din viitor s-ar extinde pentru totdeauna, dându-ne acces la o zestre cosmică nelimitată, în schimb, imaginea din centrul figurii arată că un univers în expansiune cu energie întunecată (aşa cum pare să fie universul nostru) deformează conurile luminoase făcându-le să semene cu un pahar de şampanie, şi limitând pe veci la circa 10

miliarde numărul galaxiilor pe care le putem coloniza.

Dacă această limită vă face să simţiţi claustrofobie cosmică, vreau să vă ridic moralul spunându-vă că în calculele mele am presupus că energia întunecată rămâne constantă de-a lungul timpului, ceea ce este în acord cu cele mai recente măsurători. Nu

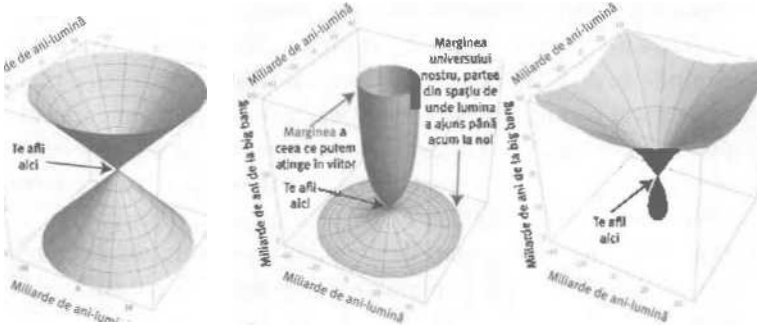


Figura 6.7Într-o diagramă spaţio-temporală, un eveniment este un punct ale cărei poziţii pe orizontală şi pe verticală reprezintă unde şi, respectiv, când are ioc. Dacă spaţiul nu se extinde (imaginea din stânga), două conuri delimitează regiunea din spaţiu-timp care pe noi, aflaţi pe Pământ (vârful conurilor), ne poate influenţa (conul de jos), şi cea pe care o putem influenţa (conul de sus). Lucrurile devin mai interesante atunci când spaţiul se extinde (imaginile din centru şi dreapta). Conform modelului standard din cosmologie. putem vedea şi atinge doar o parte finită a spaţiului-timp, chiar dacă spaţiul e infinit. în imaginea din mjiloc, semănând cu un pahar de şampanie, folosim coordonate care ascund expansiunea spaţiului, aşa încât mişcările galaxiilor în decursul timpului corespund dreptelor verticale. Din punctul nostru actual de observaţie, la 13,8 miliarde de ani după big bang, au avut timp să alungă la noi doar razele de lumină de la baza paharului de şampanie, şi, chiar dacă am călători cu viteza luminii, n-am putea niciodată ajunge la regiunile din afara părţii superioare a paharului, care conţine aproximativ 10 miliarde de galaxii. în imaginea din dreapta, care seamănă cu o picătură de apă aflată sub o floare, folosim coordonatele obişnuite în care se vede expansiunea spaţiului. Aceasta deformează baza paharului făcând-o să semene cu o picătură, pentru că regiunile de la marginea a ceea ce putem vedea erau la început foarte aproape una de alta.

ştim însă deocamdată ce este de fapt energia întunecată, aşa că există o mică speranţă ca aceasta să dispară în cele din urmă (la fel cum a dispărut substanţa asemănătoare energiei întunecate a cărei existenţă a fost postulată pentru a explica inflaţia cosmică), iar dacă aşa se va întâmpla, acceleraţia va fi înlocuită de *decelerație,* permiţând în principiu viitoarelor forme de viaţă să continue să colonizeze mereu alte galaxii atâta vreme cât

vor dăinui.

Cât de repede ne putem deplasa?

Am văzut mai sus câte galaxii poate coloniza o civilizaţie dacă se răspândeşte în toate direcţiile cu viteza luminii. Teoria relativităţii ne spune că e imposibil să trimiţi rachete prin spaţiu cu viteza luminii, fiindcă ai avea nevoie de energie infinită – prin urmare, cât de repede se pot deplasa practic rachetele?††††††††††

Când a fost lansată spre Pluto în 2006, racheta NAŞA New Horizons a bătut recordul de viteză, ajungând la 45 km/s, iar sonda spaţială Parker urma să atingă în 2018 o viteză de patru ori mai mare în apropierea Soarelui, însă chiar şi această valoare reprezintă doar 0,1% din viteza luminii. Proiectarea unor rachete mai rapide a captivat unele dintre minţile strălucite ale secolului trecut, şi există o literatură bogată şi fascinantă pe această temă. De ce e atât de greu să te deplasezi mai repede? Există două mari probleme: rachetele convenţionale cheltuiesc cea mai mare parte a combustibilului pur şi simplu pentru a accelera combustibilul pe care îl transportă, iar combustibilul de azi e îngrozitor de ineficient – fracţiunea din masa Iui transformată în energie nu e mult mai mare decât acea fracţiune de 0,00000005% pentru benzină din tabelul 6.1.0 îmbunătăţire evidentă ar fi trecerea la un combustibil mai eficient. De exemplu, o echipă din care facea parte şi Freeman Dyson a lucrat la Proiectul Orion al NAŞA, în care, prin detonarea a 300 000 de bombe nucleare în cursul a zece zile, o navă suficient de mare pentru a transporta oameni într-o călătorie de un secol spre alt sistem solar să fie accelerată până la 3% din

†††††††††† Un calcul simplu arată că, dacă civilizaţia se răspândeşte prin spaţiu nu cu viteza luminii c, ci o cu o viteză v mai mică, numărul galaxiilor colonizate scade cu un factor $(t'/c)^3$. Asta înseamnă că o civilizaţie mai leneşă, care se răspândeşte cu o viteză de 1 o ori mai mică decât cea a luminii, va coloniza de 1000 de ori mai puţine galaxii. (N. 4)

viteza luminii.[5] Altă idee presupunea folosirea antimateriei drept combustibil, fiindcă prin combinarea acesteia cu materia obişnuită se eliberează energie cu o eficienţă de aproape 100%.

Altă propunere este de a construi o rachetă care nu-şi transportă combustibilul. De exemplu, spaţiul interstelar nu e complet vid, ci conţine cantităţi mici de ioni de hidrogen (atomi de hidrogen care şi au pierdut electronul, adică protoni). Asta i-a sugerat în 1960 fizicianului Robert Bussard ideea a ceea ce numim acum *colector Bussard:* în traiectoria ei, racheta culege aceşti ioni şi îi foloseşte drept combustibil într-un reactor de fuziune aflat la bord. Cercetări recente arată că punerea în practică a ideii e îndoielnică, dar există o altă propunere de navigare spaţială fără a transporta combustibil, care pare să fie realistă pentru o civilizaţie cu tehnologie avansată: navigarea cu ajutorul laserului.

Figura 6.8 ilustrează un proiect de rachetă propulsată prin laser, conceput în 1984 de Robert Forward, acelaşi care a propus construirea unei sfere Dyson monolitice staţionare. La fel cum moleculele de aer ricoşând pe vela unei ambarcaţiuni o împing, particulele de lumină (fotonii) ricoşând pe o oglindă împing racheta. Orientând fasciculul unui laser uriaş alimentat cu energie solară spre o imensă velă foarte uşoară ataşată de navă, putem folosi energia Soarelui pentru a accelera racheta până la viteze mari. Dar cum o oprim? Am aflat răspunsul din strălucitul articol al lui Forward: după cum se vede în figura 6.8, inelul exterior al velei se detaşează şi e deplasat în faţa navei, reflectând fasciculul laser pentru a decelera nava şi vela mai mică.[6] Forward a calculat că în felul acesta am putea face călătoria de patru ani-lumină până la sistemul solar a Centauri în doar 40 de ani. Ajunşi acolo, am construi un alt laser gigant şi ne am continua din stea în stea călătoria prin Calea Lactee.

De ce să ne oprim însă aici? în 1964, astronomul sovietic Nikolai Kardaşev a propus clasificarea civilizaţiilor după cantitatea de energie pe care o pot folosi. Exploatarea energiei unei planete, a unei stele (cu o sferă Dyson, de pildă) şi a unei galaxii corespunde civilizaţiilor de Tipul I, Tipul II şi, respectiv, Tipul III pe scara lui Kardaşev. S-a propus ulterior ca Tipul IV să corespundă exploatării întregului nostru univers accesibil. între timp au apărut o veste bună şi o veste proastă pentru formele de viaţă ambiţioase. Vestea proastă este că există energia întunecată, care, după cum am văzut, pare să ne limiteze raza de acţiune. Vestea bună este că

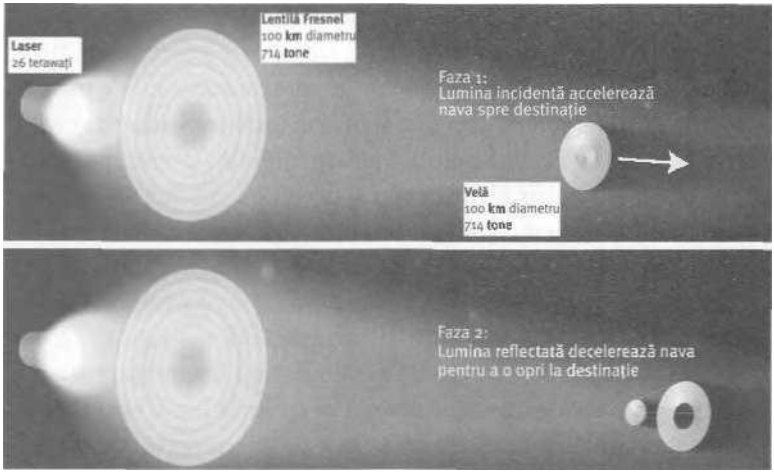


Figura 6.8 Proiectul lui Robert Forward pentru navigare cu ajutorul laserului spre steaua a Centauri, aflată la distanţa de patru ani-lumină. Iniţial, un laser puternic din sistemul nostru solar accelerează nava aplicând presiunea radiaţiei asupra velei. Pentru a frâna înainte de a ajunge la destinaţie, partea exterioară a velei e detaşată şi reflectă fasciculul laser spre navă.

inteligenţa artificială a făcut progrese spectaculoase. Până şi vizionari optimişti precum Cari Sagan credeau că sunt şanse mici ca oamenii să ajungă în alte galaxii, dat fiind că astronauţii ar muri în prima sută de ani a unei călătorii care ar dura milioane de ani, chiar dacă viteza s-ar apropia de cea a luminii. Refuzând să cedeze, ei se gândeau Ia congelarea astronauţilor pentru a le prelungi viaţa, la încetinirea îmbătrânirii prin deplasarea cu viteză mare sau la trimiterea unei comunităţi care ar călători vreme de zeci de mii de generaţii – mai mult decât a supravieţuit până acum specia umană.

Posibilitatea superinteligenţei schimbă complet datele problemei, făcând să crească optimismul amatorilor de călătorii intergalactice. înlăturând nevoia de a transporta sisteme masive care să susţină viaţa de tip uman şi adăugând tehnologia inventată de IA, colonizarea intergalactică încetează să mai fie un vis. Navigarea cu ajutorul laserului devine mult mai ieftină dacă racheta nu

conţine decât o „sondă-germene“; un robot capabil să se aşeze pe un

asteroid sau pe o planetă din sistemul solar vizat şi să construiască o nouă civilizaţie. Nu e nevoie nici să transporte instrucţiuni, trebuie doar să construiască o antenă receptoare suficient de mare pentru a primi planuri detaliate şi instrucţiuni transmise de civi- lizaţia-mamă cu viteza luminii. Apoi, îşi foloseşte laserele pe care le-a construit pentru a trimite noi sonde-germene care să colonizeze simultan sistemele solare ale galaxiei. Chiar şi vastele spaţii intergalactice conţin destule stele (care şi-au părăsit galaxiile) ce pot fi folosite pentru escale, făcând astfel posibilă o strategie a „saltului de la o insula Ia alta" pentru navigarea intergalactică.

Odată ce IA superinteligentă a colonizat un sistem solar sau o galaxie, aducerea oamenilor acolo n-ar fi o problemă – presupunând că oamenii au reuşit să facă din asta obiectivul IA. întreaga informaţie necesară despre oameni poate fi transmisă cu viteza luminii, după care IA ar asambla cuarci şi electroni în oamenii doriţi. Aceasta s ar putea face fie folosind o tehnologie destul de pri- mitivă, transmiţând pur şi simplu cei doi gigabyţi de informaţie necesari pentru specificarea ADN-ului unei persoane şi incubând apoi un prunc care să fie crescut de IA, fie prin nanoasamblarea de către IA din cuarci şi electroni a unui adult care ar avea scanate toate amintirile originalului de pe Pământ.

Asta înseamnă că, presupunând că are loc o explozie a inteligenţei, întrebarea-cheie nu este dacă e posibilă colonizarea intergalactică, ci cât de repede se poate ea desfăşura. Cum toate ideile prezentate aici provin de la oameni, ele trebuie privite doar ca limite inferioare ale vitezei de expansiune a vieţii; viaţa superinteligentă ambiţioasă s-ar descurca pesemne mult mai bine şi ar fi puternic stimulată să împingă şi mai departe limitele, căci în cursa contra timpului şi a energiei întunecate fiecare creştere de a vitezei medii de colonizare

înseamnă cu 3% mai multe galaxii colonizate.

De exemplu, dacă durează 20 de ani să parcurgi 10 ani-lu- mină până la următorul sistem solar folosind metoda laser-velă, iar aiţi 10 ani ca să-l colonizezi şi să construieşti acolo noi lasere şi sonde-germene, regiunea colonizată din spaţiu va fi o sferă dilatânduse în toate direcţiile în medie cu o treime din viteza luminii. Intr-o analiză aprofundată din 2014 privind civilizaţiile care se extind la scară cosmică, fizicianul american Jay Olson propunea o alternativă la „saltul de pe o insulă pe alta“, implicând două tipuri de sonde: *sonde-germene* şi *expandoare.*[7] Sondele-germene ar încetini, s-ar aşeza pe destinaţiile lor şi ar însămânţa acolo viaţa. Expandoarele însă nu s-ar opri niciodată: ar aduna materie în cursul zborului, folosind pesemne o variantă perfecţionată a tehnologiei colectorului, şi ar folosi materia drept combustibil şi pentru a construi sonde-germene şi copii ale lor însele. Această flotă auto reproducătoare de expandoare ar accelera uşor pentru a păstra mereu o viteză constantă (să zicem, jumătate din viteza luminii) în raport cu galaxiile învecinate, şi s-ar reproduce suficient de des pentru ca flota să formeze o coajă sferică în expansiune, cu un număr constant de expandoare pe aria cojii.

în fine, dar nu în ultimul rând, există ingenioasa abordare „Ave Maria“ pentru o expansiune şi mai rapidă: folosind trucul „spa mului cosmic“ al lui Hans Moravec din capitolul 4. Trimiţând un mesaj care păcăleşte civilizaţiile naive proaspăt evoluate pentru a construi o maşinărie superinteligentă care le jefuieşte, o civilizaţie se poate extinde practic cu viteza luminii, viteza cu care seducătorul ei cântec de sirenă se răspândeşte prin cosmos. Din moment ce aceasta ar putea fi *singura* soluţie pentru civilizaţiile avansate de a ajunge Ia majoritatea galaxiilor din conul luminos al viitorului, iar ele n-ar prea avea motive să n-o aplice, ar trebui să fim

foarte precauţi cu mesajele primite de la extratereştri! în cartea lui Cari Sagan *Contact,* noi, pământenii, folosim planuri primite de Ia extratereştri pentru a construi o maşinărie pe care n-o înţelegem – nu recomand să facem asta...

în concluzie, cred că majoritatea oamenilor de ştiinţă şi a autorilor SF au abordat cu pesimism problema colonizării cosmosului, ignorând posibilitatea apariţiei unei superinteligenţe: concen- trându-şi atenţia doar asupra călătorilor umani, ei au supraestimat dificultatea călătoriei intergalactice, iar concentrându-şi atenţia doar asupra tehnologiei inventate de oameni, au supraestimat timpul necesar pentru a atinge limitele fizice a ceea ce este posibil.

Păstrarea comunicării prin inginerie cosmică

Dacă energia întunecată continuă să accelereze îndepărtarea galaxiilor una de alta, aşa cum sugerează cele mai recente observaţii, vor apărea mari dificultăţi pentru viitorul vieţii. Chiar dacă o civilizaţie viitoare reuşeşte să colonizeze un milion de galaxii, în cursul a zeci de miliarde de ani energia întunecată va fragmenta acest imperiu cosmic în mii de regiuni incapabile să comunice între ele. Dacă viaţa viitoare nu face nimic pentru a împiedica această fragmentare, atunci cele mai mari bastioane ale vieţii care vor rezista vor fi grupări conţinând vreo mie de galaxii a căror gravitaţie combinată e suficient de puternică pentru a echilibra energia întunecată care încearcă să le separe.

Dacă o civilizaţie superinteligentă vrea să rămână conectată, va trebui să întreprindă lucrări de inginerie cosmică la scară mare. Cât de multă materie va avea timp să adune în cea mai mare aglomerare a ei, înainte ca energia întunecată s-o plaseze pentru totdeauna dincolo de raza ei de acţiune? O metodă de a deplasa o stea pe distanţe mari este aceea de a împinge o a treia stea într-un sistem binar în care două stele se rotesc

stabil una în jurul celeilalte. La fel ca în relaţiile amoroase, introducerea unui al treilea partener poate avea un efect destabilizator, unul dintre cei trei parteneri fiind violent respins – în cazul stelelor, cu viteză mare. Dacă unii dintre parteneri sunt găuri negre, o asemenea tripletă instabilă poate fi folosită pentru a azvârli materie în afara galaxi- ei-gazdă. Din păcate, prin această tehnică a celor trei corpuri, fie că e aplicată stelelor, găurilor negre sau galaxiilor, s-ar părea că nu poate fi deplasată decât o infimă parte din masa civilizaţiei pe distanţele mari necesare pentru a învinge energia întunecată.

Asta nu înseamnă însă că viaţa superinteligentă nu poate găsi soluţii mai bune – de pildă, transformând o mare parte din masa galaxiilor îndepărtate în nave spaţiale care pot călători spre aglomerarea centrală. Dacă poate fi construit un sfalerizor, pesemne că acesta ar putea fi folosit pentru a converti materia în energie care să fie trimisă spre aglomerarea centrală sub formă de lumină, iar acolo lumina să fie reconfigurată în materie sau utilizată ca atare.

Cea mai mare şansă ar fi dacă am reuşi să construim găuri de vierme *(wormholes)* stabile traversabile, permiţând comunicarea şi călătoria aproape instantanee între cele două capete ale găurii de vierme, indiferent cât de mare e distanţa dintre ele. Gaura de vierme este o „scurtătură“ în spaţiu-timp prin care ajungi din punctul A în punctul B fără a străbate spaţiul dintre cele două puncte. Deşi găurile de vierme stabile nu sunt interzise de relativitatea generală şi au apărut în filme precum *Contact şi Interstellar,* ele presupun existenţa unui ipotetic tip de materie cu densitate negativă, care depinde de efecte ale gravitaţiei cuantice încă necunoscute. Cu alte cuvinte, s-ar putea dovedi că găurile de vierme utile sunt imposibile – în caz contrar însă, viaţa superinteligentă are toate motivele să le construiască. Găurile de vierme nu numai că

revoluţionează comunicarea rapidă între galaxii, dar legând de timpuriu galaxiile îndepărtate de aglomerarea centrală, găurile de vierme ar permite domeniului vieţii viitoare să rămână conectat un timp nedefinit, anulând complet încercările energiei întunecate de a cenzura comunicarea. Odată ce două galaxii sunt legate printr-o gaură de vierme stabilă, ele rămân legate oricât s-ar îndepărta una de alta.

Dacă, în ciuda eforturilor de inginerie cosmică, o civilizaţie a viitorului conchide că e condamnată să piardă pe veci legătura cu zone ale ei, s-ar putea pur şi simplu despărţi de ele, urându-le succes. Dacă însă are obiective de calcul ambiţioase, implicând căutarea răspunsurilor la întrebări extrem de dificile, ar putea apela la stra tegia „pământului pârjolit“: ar converti galaxiile îndepărtate în calculatoare masive care transformă materia şi energia lor m calcule efectuate într-un ritm frenetic, cu speranţa că, înainte ca energia întunecată să împingă resturile lor epuizate dincolo de orizont, acestea ar putea transmite răspunsurile lor îndelung căutate către aglomerarea centrală. Această strategie a „pământului pârjolit“ ar fi potrivită mai cu seamă regiunilor atât de îndepărtate, încât se poate ajunge la ele doar prin metoda „spamului cosmic“, spre disperarea locuitorilor lor. în schimb, în aglomerarea centrală, civilizaţia s-ar strădui ca prin maximă conservare şi eficienţă să reziste cât mai mult cu putinţă.

Cât de mult poţi dăinui?

Majoritatea persoanelor, organizaţiilor şi naţiunilor ambiţioase aspiră la longevitate. Prin urmare, dacă o civilizaţie ambiţioasă a viitorului dezvoltă superinteligenţa şi vrea longevitate, cât de mult poate dăinui?

Prima analiză ştiinţifică aprofundată a viitorului nostru îndepărtat a fost efectuată de nimeni altul decât Freeman Dyson, iar tabelul 6.3 rezumă câteva dintre rezultatele sale importante. Concluzia este că, fără

intervenţia inteligenţei, sistemele solare şi galaxiile sunt treptat distruse, urmând apoi tot restul şi rămân doar spaţiul vid, rece şi mort, cu o strălucire a radiaţiei stingân- du-se în eternitate. Dar Freeman îşi încheie analiza pe un ton optimist: „Există motive ştiinţifice întemeiate pentru a vorbi cu seriozitate despre posibilitatea ca viaţa şi inteligenţa să modeleze acest univers al nostru conform scopurilor lor.“[8]

Ce	**Când**
Vârsta actuală a universului	io^{10} ani
Energia întunecată împinge majoritatea galaxiilor, făcându-le inaccesibile	io^{11} ani
Ultimele stele au ars complet	10^{14} ani
Planetele se detaşează de stele	to’S ani
Stelele se detaşează de galaxii	10^{19} ani
Destabilizarea orbitelor prin radiaţie gravitaţională	io^{20} ani
Dezintegrarea protonilor (cel mai curând)	> io^{34} ani
Evaporarea găurilor negre cu mase cât ale stelelor	10^{67} ani
Evaporarea găurilor negre supermasive	10^{91} ani
întreaga materie se dezintegrează transformându-se în fier	10^{1300} ani
întreaga materie formează găuri negre, care apoi se evaporă	io«»’^{4}

Tabelul 6.3 Estimări privind viitorul îndepărtat aparţinând (cu excepţia celei de-a doua şi celei de-a şaptea) lui Freeman Dyson. El a efectuat aceste calcule înainte de descoperirea energiei întunecate, care ar putea provoca anumite tipuri de „cosmocalipse“ peste io^{10}-io” ani. Protonii ar putea fi absolut stabili; în caz contrar, experimentele sugerează că vor trece peste 10^{34} ani înainte ca jumătate dintre ei să se dezintegreze.

Cred că superinteligenţa ar rezolva lesne multe dintre problemele enumerate în tabelul 6.3, pentru că poate rearanja materia în ceva mai bun decât sisteme solare şi galaxii. Provocări deseori discutate, cum e moartea Soarelui nostru peste câteva miliarde de ani, nu ne vor opri din drum, fiindcă până şi o civilizaţie cu un nivel tehnologic relativ modest se poate cu uşurinţă muta pe stele cu masă

mică, al căror timp de viaţă depăşeşte 200 de miliarde de ani. Presupunând că civilizaţiile superinteligente îşi construiesc propriile lor centrale energetice mai eficiente decât stelele, ele şi-ar dori să *împiedice* formarea stelelor pentru a conserva energia: chiar dacă ar folosi o sferă Dyson pentru a colecta întreaga energie produsă de o stea în timpul ei de viaţă principal (obţinând aproximativ 0,1% din energia totală înmagazinată în stea), s-ar putea să nu recupereze prea mult din restul de 99,9% atunci când o stea foarte masivă moare. O stea masivă sfârşeşte printr-o supernovă, o explozie în care majoritatea energiei e disipată sub formă de neutrini, particule greu de detectat, iar la stelele foarte masive o mare cantitate de masă e irosită în formarea unei găuri negre din care energia se scurge în ro[67] ani.

Câtă vreme viaţa superinteligentă nu şi-a epuizat materia/ener- gia, ea îşi poate menţine habitatul în starea pe care şi-o doreşte. Pesemne că ar putea descoperi un mijloc de a împiedica protonii să se dezintegreze folosind aşa-numitul efect *watched-pot*‡‡‡‡‡‡‡‡ din mecanica cuantică, prin care procesul de dezintegrare e încetinit dacă faci observaţii în mod regulat. Există însă riscul ca peste 10-100 de miliarde de ani să se producă o *cosmocalipsă* care să distrugă întregul nostru univers. Descoperirea energiei întunecate şi cercetări din teoria corzilor conduc la noi scenarii ale cosmocalipsei pe care Freeman Dyson nu le cunoştea când şi-a scris articolul.

Cum va sfârşi aşadar universul nostru peste miliarde de ani? Există cinci scenarii principale pentru apocalipsa cosmică, sau cosmocalipsa, ilustrate în figura 6.9: *marele îngheţ(big chill), marea implozie (big crunch), marea*

‡‡‡‡‡‡‡‡‡ Expresia derivă din proverbul englezesc «A watched pot never boils“ („Oala pe care o păzeşti nu fierbe“). Numit şi efect Zenon cuantic, acesta se bazează pe faptul că o măsurătoare efectuată asupra unui sistem cuantic influenţează starea sistemului, deci măsurători frecvent repetate ar putea „îngheţa“ sistemul intr-o anumită stare, blocându-i evoluţia. *(N.* t)

ruptură (big rip), marea plesnitură (big snap) şi *bule mortale(death bubbles).* Universul nostru se extinde de circa 14 miliarde de ani. Marele îngheţ presupune ca universul să se extindă pe veci, diluând cosmosul într-un spaţiu rece, întunecos şi, în cele din urmă, mort; acesta era considerat rezultatul final cel mai probabil atunci când Freeman şi-a scris articolul. îl consider alegerea lui T.S. Eliot; ^Aşa sfârşeşte lumea / Nu în explozie, ci în scâncet.“ Dacă, asemenea lui Robert Frost, aţi vrea ca lumea să sfârşească în flăcări, nu în gheaţă, rugaţi-vă să aibă loc marea implozie, caz în care expansiunea cosmică e în cele din urmă inversată, iar totul se prăbuşeşte într un colaps catastrofal, ca un big bang derulat în sens invers. Marea ruptură seamănă oarecum cu marele îngheţ, galaxiile, planetele şi atomii se dezmembrează într-un final grandios peste un timp finit. Pe care dintre aceste trei scenarii ai paria? Asta depinde de ce se întâmplă cu energia întunecată, care reprezintă aproximativ 70% din masa universului nostru, pe măsură ce spaţiul continuă să se extindă. Poate fi oricare dintre scenariile îngheţ, implozie sau ruptură, după cum energia întunecată rămâne neschimbată, se diluează până la o densitate negativă sau, respectiv, se antidiluează până la o densitate mai mare. Fiindcă deocamdată n-avem nici un indiciu privind energia întunecată, iată cum aş paria eu: 40% pe marele îngheţ, 9% pe marea implozie şi 1% pe marea ruptură.

Dar ce-aş face cu restul de 50% din banii mei? îi economisesc pentru o posibilitate care nu e nici una dintre acestea trei, fiindcă ar trebui să recunoaştem că există lucruri fundamentale pe care deocamdată nu le înţelegem. Natura spaţiului, de pildă. Sfârşitul prin îngheţ, implozie sau ruptură presupune că spaţiul însuşi e stabil şi poate fi întins oricât de mult. Pe vremuri, credeam că spaţiul e doar scena inertă pe care se desfăşoară drama cosmică. Apoi Einstein ne-a învăţat că spaţiul este unul dintre actorii principali: se poate curba pentru a da naştere găurilor negre, se poate undui sub

formă de unde gravitaţionale şi se poate întinde ca univers în expansiune. Pesemne că poate chiar îngheţa într-o altă fază,

CUM VA SFÂRŞI UNIVERSUL NOSTRU?

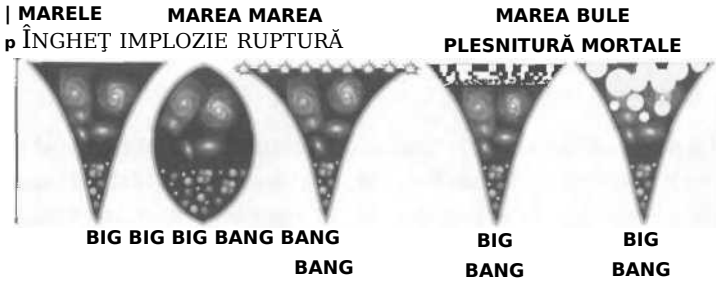


Figura 6.9 Ştim că universul nostru s-a născut dintr-un big bang fierbinte cu 14 miliarde de ani în urmă, s-a extins şi s-a răcit, a agregat particulele în atomi, stele şi galaxii. Dar nu ştim care e soarta lui. Printre scenariile propuse se numără un mare îngheţ (expansiune eternă), o mare implozie (colaps), o mare ruptură (o rată infinită de expansiune care dezmembrează totul), o mare plesnitură (textura spaţiului dezvăluind o natură granulară letală atunci când e întinsă prea mult) şi bule mortale (spaţiul „înghetând" în bule letale care se extind cu viteza luminii)

oarecum asemănător apei, iar bulele mortale ale noii faze care se extind rapid devin un nou candidat pentru cosmocalipsă. Dacă bulele mortale sunt posibile, ele s-ar extinde probabil cu viteza luminii, Ia fel ca sfera în permanentă dilatare a spamului cosmic trimis de o civilizaţie extrem de agresivă.

Teoria lui Einstein mai spune că întinderea spaţiului poate continua nelimitat, permiţând universului să ajungă la un volum infinit, ca în scenariile îngheţului şi rupturii – ceea ce bănuiesc că nu-i adevărat O bandă de cauciuc pare continuă, la fel ca spaţiul, dar dacă o întinzi prea mult, plesneşte. De ce? Fiindcă e alcătuită din atomi, iar dacă e întinsă suficient de mult, devine importantă această natură atomică granulară a benzii. Ar putea oare avea şi spaţiul o asemenea structură granulară la o scară prea mică pentru

ca noi s-o fi observat? Cercetările de gravitaţie cuantică sugerează că n-are sens să vorbim despre spaţiul tridimensional tradiţional la scări mai mici de aproximativ io^{-*4} metri. Dacă e adevărat că spaţiul nu poate fi întins la nesfârşit fără a

suferi o „mare plesnitură” catastrofală, atunci civilizaţiile viitorului şi-ar dori pesemne să se mute în cea mai mare regiune a spaţiului care nu se extinde (un roi uriaş de galaxii) la care pot ajunge.

Cât de mult poţi calcula?

După ce am văzut cât de mult *poate* dăinui viaţa, să vedem cât de mult *şi-ar dori* să dăinuie. Pare firesc să vrei să trăieşti cât mai mult, însă Freeman Dyson *a* venit şi cu un argument cantitativ pentru această dorinţă: costul calculului scade atunci când calculezi mai lent, aşadar în cele din urmă obţii mai mult dacă încetineşti lucrurile cât mai mult cu putinţă. Freeman chiar a calculat că, dacă universul nostru continuă pe veci să se extindă şi să se răcească, e posibilă o cantitate infinită de calcule.

Lent nu înseamnă neapărat plictisitor: dacă viaţa viitoare trăieşte într-o lume simulată, curgerea timpului percepută subiectiv nu are neapărat legătură cu ritmul glaciar în care simularea se derulează în lumea exterioară, deci perspectivele unui calcul infinit s-ar putea traduce printr-o nemurire subiectivă pentru formele de viaţă simulate. Cosmologul FranckTipler a pornit de aici pentru a lansa ideea că am putea ajunge la nemurirea subiectivă în momentele finale dinaintea marii implozii, făcând ca viteza de calcul să tindă spre infinit atunci când temperatura şi densitatea cresc enorm.

Din moment ce energia întunecată pare să anuleze atât speranţa lui Freeman, cât şi pe cea a lui Franck de a ajunge la calculul infinit, superinteligenţa viitorului ar putea prefera să-şi consume rezervele de energie relativ rapid, să le transforme în calcul înainte de a se lovi de dificultăţi cum sunt cele legate de orizonturile cosmice şi de dezintegrarea protonului. Dacă maximizarea cantităţii totale de calcul este obiectivul suprem, strategia optimă

va fi un compromis între prea lent (pentru a evita dificultăţile pomenite mai sus) şi prea rapid (cheltuind mai multă energie per calcul decât e nevoie).

Tot ce am discutat în acest capitol ne arată că centrale şi calculatoare de eficienţă maximă ar permite vieţii superinteligente să efectueze o cantitate extraordinară de calcule. Alimentarea creierului nostru de 13 waţi vreme de o sută de ani necesită energia din aproximativ o jumătate de miligram de materie – mai puţin decât un grăunte de zahăr. Conform studiilor lui Seth Uoyd, creierul ar putea avea o eficienţă energetică de un trilion (10^{15}) de ori mai mare, permiţând acelui grăunte de zahăr să alimenteze o simulare a tuturor vieţilor oamenilor care au trăit vreodată şi a vieţilor de încă o mie de ori mai multor oameni. Dacă toată materia din universul nostru disponibil ar fi folosită pentru a simula oameni, s-ar ajunge la peste 10^{69} vieţi – sau ce-ar prefera superinteligenţa IA să facă cu puterea ei de calcul. Şi mai multe vieţi ar fi posibile dacă simulările lor ar fi rulate mai lent.[9] Pe de altă parte, în cartea sa *Superintelligence,* Nick Bostrom estimează că, pornind de la presupuneri mai puţin optimiste privind eficienţa energetică, ar putea fi simulate io^{s8} vieţi umane. Oricât am analiza şi reanaliza cifrele, ele rămân uriaşe, Ia fel ca responsabilitatea noastră de a nu irosi acest potenţial al vieţii viitoare. „Dacă am reprezenta întreaga fericire trăită în cursul unei asemenea vieţi printr-o singură lacrimă de bucurie“, spune Bostrom, „atunci fericirea acestor suflete ar umple şi reumple oceanele Pământului în fiecare secundă vreme de o sută de miliarde de miliarde de milenii. Important este să ne asigurăm că acestea sunt cu adevărat lacrimi de bucurie.“

IERARHII COSMICE

Viteza luminii limitează nu doar răspândirea vieţii, ci şi natura ei, impunând constrângeri drastice asupra comunicării, conştiinţei şi controlului. Aşadar, dacă o mare parte din cosmosul nostru va căpăta în cele din urmă viaţă, cum va arăta această viaţă?

Ierarhii ale gândirii

Ţi s-a întâmplat vreodată să încerci să prinzi cu mâna o muscă şi să nu reuşeşti? Musca reacţionează mai repede decât tine pentru că e mai mică, aşa încât informaţiei îi ia mai puţin timp să parcurgă distanţele dintre ochi, creier şi muşchi. Acest principiu „mai mare = mai lent** se aplică nu numai în biologie, unde limita de viteză e impusă de rapiditatea cu care semnalele electrice circulă prin neuroni, dar şi în viaţa cosmică viitoare, presupunând că intr-adevăr informaţia nu se poate propaga mai repede decât lumina. Pentru un sistem inteligent de procesare a informaţiei, creşterea în dimensiuni prezintă deci şi dezavantaje. Pe de o parte, a fi mai mare înseamnă a conţine mai multă materie, ceea ce dă posibilitatea unor gânduri mai complexe. Pe de altă parte, aceasta încetineşte ritmul în care pot apărea gânduri cu adevărat globale, căci acum durează mai mult până ce informaţia relevantă ajunge la toate componentele.

Prin urmare, dacă viaţa va înghiţi cosmosul nostru, ce formă va alege: simplă şi rapidă, sau complexă şi lentă? Presupun că va face aceeaşi alegere pe care a facut-o şi viaţa de pe Pământ: ambele! Dimensiunile locuitorilor biosferei Pământului se întind pe un spectru vast, de la balena albastră cântărind două sute de tone până la bacteria de 10^{-16} kg numită *Pelagibacter,* a cărei biomasă se crede că o depăşeşte pe cea a tuturor peştilor din lume. Apoi, la organismele mari, complexe şi lente încetineala e adesea contrabalansată de modulele lor mai mici care sunt simple şi rapide. De pildă, reflexul nostru de a clipi este extrem de rapid tocmai pentru că e pus în

funcţiune de un circuit mic şi simplu care nu implică cea mai mare parte a creierului: dacă acea muscă greu de prins se îndreaptă din întâmplare spre ochiul tău, vei clipi într-o zecime de secundă, cu mult înainte ca informaţia relevantă să aibă timpul să se răspândească în creier şi să te facă să devii conştient de ce se petrece. Organizând procesarea informaţiei într-o ierarhie de module, biosfera noastră reuşeşte să obţină şi viteză, şi complexitate. Noi, oamenii, folosim deja aceeaşi strategie ierarhică pentru a optimiza calculul paralel.

Deoarece comunicarea internă e lentă şi costisitoare, mă aştept ca viaţa cosmică avansată din viitor să procedeze la fel, aşa încât calculele să fie efectuate, pe cât se poate, local. Dacă un calcul e suficient de simplu pentru a fi efectuat cu ikg de computer, e contraproductiv să-l răspândeşti într-un computer de dimensiunea unei galaxii, fiindcă aşteptarea ca informaţia să fie partajată cu viteza luminii după fiecare pas de calcul provoacă o întârziere ridicolă de circa zoo ooo de ani la fiecare pas.

Ce anume din această viitoare procesare a informaţiei va fi făcut în mod *conştient* (presupunând că asta se va întâmpla), în sensul că va implica o experienţă subiectivă, e o problemă delicată şi fascinantă pe care o vom trata în capitolul 8. Dacă a fi conştient presupune ca diferite părţi ale sistemului să poată comunica între ele, atunci gândirea sistemelor mari e în mod necesar mai lentă. în vreme ce tu sau un supercomputer viitor de dimensiunea Pământului puteţi avea multe gânduri pe secundă, o minte de dimensiunea unei galaxii poate avea doar un gând la fiecare sută de mii de ani, iar o minte cosmică atingând în dimensiune un miliard de ani-lumină ar avea răgaz doar pentru zece gânduri înainte ca energia întunecată s-o fragmenteze în părţi între care nici o legătură să nu fie posibilă. Nu-i mai puţin adevărat însă că aceste câteva gânduri preţioase şi trăirile care Ie însoţesc ar putea fi

nebănuit de profunde!

Ierarhii ale controlului

Dacă gândirea însăşi e organizată într-o ierarhie care se întinde pe un spectru larg de dimensiuni, ce putem spune despre putere? în capitolul 4 am văzut că entităţile inteligente se organizează spontan în ierarhii ale puterii ajunse Ia echilibru Nash, în care orice entitate s-ar descurca mai rău dacă şi-ar modifica strategia. Pe măsură ce se perfecţionează tehnologia transportului şi comunicării, aceste ierarhii cresc. Dacă superinteligenţa se va extinde într-o bună zi la scară cosmică, cum va arăta ierarhia puterii? Va fi descentralizată sau autoritară? Cooperarea se va baza în principal pe avantajul reciproc sau pe coerciţie şi ameninţare?

Pentru a lămuri aceste probleme, trebuie să avem în vedere atât „morcovul", cât şi „băţul": ce stimulente există pentru colaborarea la scară cosmică, şi ce ameninţări ar putea fi folosite pentru a o impune?

Controlul cu ajutorul „morcovului"

Pe Pământ, comerţul a fost un motor tradiţional al cooperării pentru că dificultatea relativă de a produce anumite lucruri variază de la un loc la altul. Dacă extragerea prin minerit a unui kilogram de argint costă de 300 de ori mai mult decât extragerea unui kilogram de cupru într-o regiune, dar numai de 100 de ori mai mult decât în alta, ambele regiuni vor avea de câştigat dacă vor schimba 200 kg de cupru pentru 1 kg de argint. Dacă o regiune posedă o tehnologie mult mai avansată decât o alta, ambele ar avea de câştigat schimbând bunuri de înaltă tehnologie contra materii prime.

Dacă însă superinteligenţa dezvoltă o tehnologie prin care poate cu uşurinţă rearanja particulele în orice formă de materie, aceasta va descuraja în mare parte

comerţul pe distanţe lungi. De ce să te chinui să transporţi argint între două sisteme solare îndepărtate, când e mai simplu şi mai rapid să transformi cuprul în argint rearanjându-i particulele? De ce să te chinui să transporţi dispozitive de înaltă tehnologie între galaxii, când informaţiile tehnice şi materiile prime (orice materie e bună) există în ambele locuri? Bănuiesc că, într-un cosmos înţesat cu superinteligenţă, aproape singurul bun care merită transportat pe distanţe lungi va *fi informaţia.* Unica excepţie ar putea fi materia folosită în proiectele de inginerie cosmică – de pildă, pentru a contracara tendinţa de- structivă a energiei întunecate de a separa civilizaţiile. Spre deosebire de ce se întâmplă în comerţul uman tradiţional, această materie poate fi transportată sub orice formă compactă convenabilă, eventual chiar ca fascicul de energie, din moment ce superinteligenţa care o primeşte o poate rapid rearanja în orice obiecte vrea.

Dacă partajarea sau vinderea informaţiei devine principalul motor al cooperării cosmice, ce tipuri de informaţii ar putea fi implicate? Orice informaţie dorită va deveni preţioasă dacă generarea ei necesită un efort de calcul masiv şi îndelungat. De exemplu, o superinteligenţă ar putea dori să afle răspunsuri la întrebări ştiinţifice complicate despre natura fizică a realităţii, Ia întrebări matematice despre teoreme şi algoritmi optimi sau la întrebări inginereşti privind construirea unei tehnologii spectaculoase.

Formele de viaţă hedoniste şi-ar putea dori distracţie digitală şi experienţe simulate fabuloase, iar comerţul cosmic ar putea stimula cererea pentru o formă de criptomonedă cosmică în genul bitcoin.

Asemenea posibilităţi de schimb ar face ca informaţia să nu circule doar între entităţi comparabile, dar şi în sus şi în jos, în cadrul ierarhiilor de putere – de pildă, între noduri de dimensiunea unui sistem solar şi un centru

galactic, sau între noduri de dimensiunea unei galaxii şi un centru cosmic. Nodurile şi-ar dori asta pentru a simţi că fac parte dintr-un ansamblu mai mare, pentru a primi răspunsuri şi tehnologii la care, singure, n-au acces, şi pentru a se apăra de ameninţări externe. Ele ar putea fi atrase şi de perspectiva unei cvasi-nemuriri prin copiere: la fel cum mulţi oameni se consolează cu credinţa că sufletul lor va supravieţui după ce trupul fizic moare, o LA avansată s-ar mulţumi ca mintea şi cunoştinţele ei să supravieţuiască într-un supercalculator din centru, după ce hardware-ul ei originar şi-a epuizat rezervele de energie.

Reciproc, centrul ar vrea ca nodurile lui să-l ajute în sarcini computaţionale masive pe termen lung, de rezultatele cărora nu are urgent nevoie, aşa încât merită să aştepte răspunsul mii sau milioane de ani. Aşa cum am văzut mai sus, centrul şi-ar putea dori şi ca nodurile să-l ajute la proiecte inginereşti de anvergură, cum ar fi contracararea energiei întunecate deplasând întregi concentraţii de masă galactică. Dacă se va dovedi că găurile de vierme traversabile există şi pot fi construite, atunci principala prioritate a centrului va fi pesemne construirea unei reţele de găuri de vierme pentru a se opune energiei întunecate şi a păstra imperiul conectat pe termen nedefinit. Problema aprig discutată a obiectivelor supreme pe care le-ar putea avea o superinteligenţă va fi analizată în capitolul 7.

Controlul cu ajutorul „băţului“

Imperiile terestre îşi constrâng de regulă subordonaţii să coopereze apelând atât la „morcov“, cât şi la „băţ“. Supuşii Imperiului Roman preţuiau tehnologia, infrastructura şi protecţia care li se ofereau în schimbul cooperării lor, dar se şi temeau de repercusiunile inevitabile în caz de revoltă sau neplata dărilor. Datorită timpului lung necesar pentru a trimite trupe de la Roma în provinciile îndepărtate, intimidarea cădea în bună măsură în sarcina trupelor

locale şi a înalţilor funcţionari loiali, care aveau misiunea de a aplica imediat pedepse. Un centru superinteligent ar putea folosi strategia asemănătoare de a desfăşura o reţea de gărzi loiale pe întinsul imperiului cosmic. Cum supuşii superinteligenţi pot fi greu de controlat, cea mai simplă strategie viabilă ar fi folosirea unor gărzi IA programate să fie 100% loiale în virtutea faptului că sunt relativ stupide, urmărind doar ca toate regulile să fie respectate, în caz contrar activând automat o armă a apocalipsei.

Să presupunem, de exemplu, că centrul IA plasează o stea pitică albă în apropierea unei civilizaţii de dimensiunea unui sistem solar pe care vrea s-o controleze. Pitica albă este ceea ce rămâne după ce o stea de masă medie îşi epuizează combustibilul. Alcătuită în mare măsură din carbon, seamănă cu un diamant uriaş, şi e atât de compactă încât poate cântări mai mult decât Soarele, deşi e mai mică decât Pământul. Fizicianul indian Subrahmanyan Chandrasekhar a demonstrat că, dacă îi adaugi masă până depăşeşte *limita Chandrasekhar,* de vreo 1,4 ori masa Soarelui nostru, va exploda într-un cataclism termonuclear numit supernovă de tipul rA. Dacă IA centrală a aranjat în mod cinic ca această pitică albă să fie foarte aproape de limita Chandrasekhar, gardianul IA ar fi eficient chiar şi dacă ar fi extrem de stupid (în bună parte, tocmai datorită stupidităţii): ar putea fi programat să verifice numai dacă civilizaţia subjugată şi-a predat cota lunară de bitcoini, demonstraţii matematice sau orice alte taxe erau stipulate, iar în caz contrar să arunce suficientă masă în pitica albă pentru a declanşa supernova, distrugând complet întreaga regiune.

Civilizaţii de dimensiunea unei galaxii ar putea fi controlate în mod asemănător, plasând multe obiecte compacte pe orbite apropiate de gaura neagră uriaşă din centrul galaxiei, şi ameninţând să transforme aceste mase în gaz – de pildă, ciocnindu-le. Acest gaz ar începe apoi să hrănească gaura neagră, transformând-o într-un

quasar puternic care ar face ca o mare parte din galaxie să devină nelocuibilă.

în concluzie, există motive puternice pentru ca viaţa din viitor să coopereze Ia distanţe cosmice, dar nu putem spune dacă această cooperare se va baza în principal pe avantaje reciproce sau pe ameninţări brutale – limitele impuse de fizică par să permită ambele scenarii, aşa încât rezultatul va depinde de obiectivele şi valorile predominante. Vom discuta despre capacitatea noastră de a influenţa aceste obiective şi valori ale vieţii viitoare în capitolul 7.

Când civilizaţiile se confruntă

Am analizat până acum doar scenarii în care viaţa se extinde în cosmos pornind de la o unică explozie a inteligenţei. Ce se întâmplă însă dacă viaţa evoluează independent în mai multe locuri, iar două civilizaţii în expansiune se întâlnesc?

Pentru un sistem solar oarecare, probabilitatea ca viaţa să evolueze pe una dintre planetele lui, să dezvolte o tehnologie avansată şi să se extindă în spaţiu pare să fie nenulă, odată ce viaţa tehnologică a evoluat aici, în sistemul nostru solar, iar legile fizicii nu interzic colonizarea cosmosului. Dacă spaţiul e suficient de vast (teoria inflaţiei cosmologice sugerează că e imens sau infinit), atunci există pesemne numeroase civilizaţii care se extind, după cum e ilustrat în figura 6.10. Articolul Iui Jay Olson menţionat mai sus conţine o elegantă analiză a unor asemenea biosfere cosmice în expansiune, iar Toby Ord, împreună cu colegii lui de la Institutul pentru Viitorul Omenirii, a efectuat o analiză similară. Privite în cele trei dimensiuni spaţiale, aceste biosfere cosmice sunt practic sfere, dacă presupunem că civilizaţiile se extind cu aceeaşi viteză în toate direcţiile. în spaţiu-timp, biosferele arată ca partea superioară a „paharului de şampanie“ din figura

6.7, fiindcă energia întunecată limitează numărul galaxiilor la care fiecare civilizaţie poate ajunge.

Dacă distanţa dintre civilizaţiile colonizatoare vecine e mult mai mare decât le permite energia întunecată să se extindă, ele nu vor intra niciodată în contact, şi nici măcar nu vor afla una de existenţa celeilalte, aşa încât fiecare va simţi că e singură în cosmos.

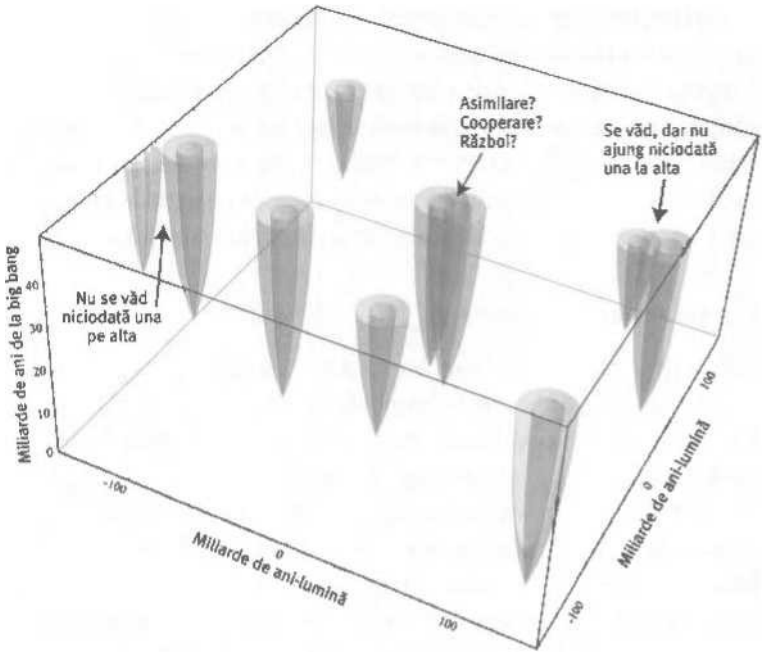


Figura 6.10 Dacă viaţa evoluează independentîn mai multe puncte din spa- ţiu-timp (locuri şi momente) şi începe să colonizeze spaţiul, atunci spaţiul va conţine o reţea de biosfere cosmice în expansiune, fiecare semănând cu partea de sus a „paharului de şampanie" din figura 6.7. Capătul de jos al fiecărei biosfere reprezintă locul şi momentul în care a început colonizarea. „Paharele de şampanie" opac şi traslucid corespund colonizării cu viteza de 50% şi, respectiv, 100% din viteza luminii, iar suprapunerile indică unde se întâlnesc civilizaţiile independente.

Dacă însă cosmosul nostru e atât de fecund încât distanţele dintre civilizaţiile vecine să fie mai mici, atunci unele civilizaţii se vor suprapune în cele din urmă. Ce se va întâmpla în aceste regiuni de suprapunere? Va fi cooperare, competiţie sau războaie?

Europenii au putut cuceri Africa şi Americile fiindcă aveau o tehnologie superioară. Dimpotrivă, e plauzibil ca, mult înainte de întâlnirea a două civilizaţii superinteligente, tehnologiile lor să se fi stabilizat la acelaşi nivel, limitat doar de legile fizicii. Prin urmare,

pare puţin probabil ca o superinteligenţă s-o cucerească lesne pe cealaltă, chiar dacă asta şi-ar dori. Apoi, dacă obiectivele lor au evoluat aproximativ în aceeaşi direcţie, nu prea ar avea motive să-şi dorească cucerirea sau războiul. De pildă, dacă ambele ar încerca să demonstreze cât mai multe teoreme frumoase şi să inventeze algoritmi cât mai ingenioşi, ele îşi pot pur şi simplu împărtăşi descoperirile în beneficiul ambelor. în definitiv, informaţia e foarte diferită de resursele pentru care oamenii se luptă de obicei, odată ce o poţi dărui, şi în acelaşi timp păstra.

Unele civilizaţii în expansiune ar putea avea obiective esenţi- almente inflexibile, ca acelea ale unui cult fundamentalist sau ale unui virus care se răspândeşte. E însă de asemenea plauzibil ca unele civilizaţii avansate să semene mai curând cu oamenii lipsiţi de prejudecăţi – gata să-şi ajusteze obiectivele când li se prezintă argumente suficient de convingătoare. Daca două dintre acestea se întâlnesc, va fi o confruntare nu de arme, ci de idei, în care triumfă cea mai persuasivă, iar obiectivele ei se răspândesc cu viteza luminii în întreaga regiune controlată de cealaltă civilizaţie. Asimilarea vecinilor e o strategie de expansiune mai rapidă decât colonizarea, fiindcă sfera ta de influenţă se poate extinde cu viteza cu care se propagă ideile (viteza luminii, folosind telecomunicaţia), în vreme ce colonizarea fizică avansează inevitabil mai lent. Această asimilare nu va fi forţată, ca aceea aplicată de Borg din *Star Trek,* ci voluntară, bazată pe superioritatea persuasivă a ideilor, cel asimilat având de câştigat.

Am văzut că universul viitor poate conţine două tipuri de bule în expansiune rapidă: civilizaţii care se extind şi acele bule mortale care se extind cu viteza luminii şi fac ca spaţiul să devină de nelocuit, distrugând toate particulele elementare. O civilizaţie ambiţioasă poate deci întâlni trei tipuri de regiuni: zone nelocuite, bule ale vieţii şi bule ale

morţii. Dacă se teme de civilizaţii rivale necooperante, are toate motivele să se lanseze într-o „acaparare a terenurilor“ şi să colonizeze zonele nelocuite înainte s-o facă rivalii. Are însă aceeaşi pornire expansionistă chiar şi dacă nu există alte civilizaţii, pur şi simplu pentru a dobândi resurse înainte ca energia întunecată să le facă să devină de neatins. Ca atare, a da peste o altă civilizaţie în expansiune poate fi un lucru mai bun sau mai rău decât a da peste un spaţiu nelocuit, în funcţie de măsura în care vecinul e cooperant şi lipsit de prejudecăţi. Oricum, e mai bine să te ciocneşti de o civilizaţie expansionistă (chiar şi de una care încearcă să te transforme în agrafe de birou) decât să te ciocneşti de o bulă mortală, care va continua să se extindă cu viteza luminii, indiferent dacă tu încerci să te lupţi sau sa porţi o discuţie rezonabilă cu ea. Singura protecţie împotriva bulelor mortale ne-o poate oferi energia întunecată, care le împiedică pe cele aflate la mare distanţă să ajungă la noi. Prin urmare, dacă bulele mortale sunt într-adevăr răspândite prin univers, energia întunecată este de fapt prietenul, nu duşmanul nostru.

Suntem oare singuri?

Mulţi oameni sunt convinşi că există viaţă avansată în numeroase locuri din univers, aşa încât extincţia omenirii n-ar însemna mare lucru din perspectivă cosmică. în fond, de ce să ne îngrijorăm că ne-am putea autodistruge, dacă o civilizaţie cu obiective nobile gen *Star Trek* şi-ar face curând apariţia şi ar reînsămânţa cu viaţă sistemul nostru solar? Cred că această ipoteză *Star Trek* e periculoasă, fiindcă ne dă o amăgitoare senzaţie de securitate, iar civilizaţia noastră riscă să devină indiferentă şi nesăbuită. Şi, de fapt, e nu doar periculoasă, dar şi probabil falsă.

Nu multă lume împărtăşeşte punctul meu de vedere,* şi s-ar putea să mă înşel, însă e o posibilitate pe care deocamdată n-o putem respinge şi care ne obligă

moralmente să facem toate eforturile pentru a evita distrugerea civilizaţiei noastre.

Când ţin o conferinţă de cosmologie, îi rog adesea pe cei din public să ridice mâinile dacă cred că există viaţă inteligentă în alt loc din universul nostru (regiunea spaţiului de unde lumina a

♦ Totuşi, în cartea sa din 2011 *Atone in the Universe,* John Gribbin ajunge la aceeaşi concluzie. Pentru spectrul larg ai perspectivelor în această problemă, recomand de asemenea cartea lui Paul Davis din 2011 *The Eerie Silence.* (JV. o.)

ajuns la noi în cei 13,8 miliarde de ani scurşi de la big bang). De fiecare dată ridică mâna aproape toată lumea, de Ia copii de grădiniţă la studenţi. Când întreb de ce, în general răspunsul este că universul nostru e atât de mare, încât undeva trebuie să existe viaţă, cel puţin judecând statistic. Să analizăm cu atenţie acest argument, identificând punctele lui slabe.

Totul ţine de un număr: distanţa tipică dintre două civilizaţii vecine din figura 6.10. Dacă e mai mare de 20 de ani-Iumină, este de aşteptat să fim singuri în universul nostru (repet, partea din spaţiu de unde lumina a ajuns la noi în cei 13,8 miliarde de ani scurşi de la big bang) şi să nu intrăm niciodată în contact cu extra- tereştri. Aşadar, cât de mare ar putea fi această distanţă? Nu prea avem indicii. încercând o estimare, distanţa până la vecina noastră este de aproximativ 1000... 000 metri, unde numărul total de zerouri ar putea fi 21, 22,23,..., 100,101,102 sau mai mare-dar probabil nu mai mic de 21, fiindcă n-avem deocamdată vreo dovadă convingătoare privind existenţa extratereştrilor(vezi figura 6.11).

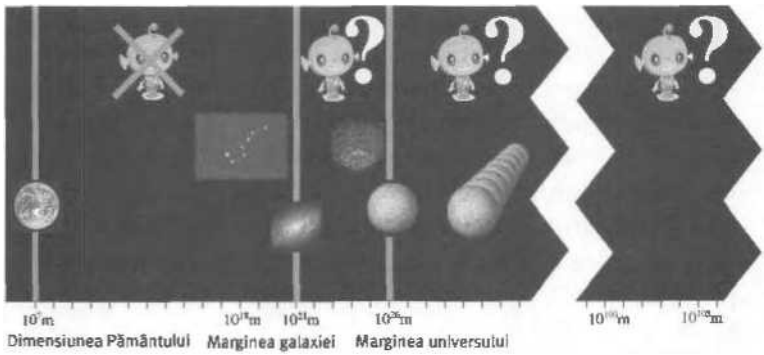


Figura 6.11 Suntem oare singuri? Uriaşa incertitudine privind felul în care evoluează viaţa şi inteligenţa ne sugerează că civilizaţia vecină cea mai apropiată se poate afla oriunde de-a lungul axei orizontale de mai sus, aşa încât este improbabil ca ea să se găsească în intervalul îngust dintre marginea galaxiei noastre (la circa 10^{21} metri de noi) şi marginea universului nostru (la circa 10^{26} metri de noi). Dacă s-ar afla la o distanţă mult mai mică, ar exista pesemne numeroase alte civilizaţii avansate în galaxia noastră, pe care probabil le-am fi observat, ceea ce sugerează că de fapt suntem singuri în univers.

Pentru ca vecina cea mai apropiată să se afle în universul nostru, a cărui rază este de circa 10^{26} metri, numărul de zerouri nu poate depăşi 26, iar probabilitatea ca numărul de zerouri să se afle în intervalul îngust dintre 22 şi 26 e destul de mică. De aceea cred că suntem singuri în univers.

Prezint o justificare amănunţită a acestui argument în cartea mea *OurMathematical Universe,* aşa încât n-am s-o reiau aici, însă motivul principal pentru care n-avem indicii privind această distanţă până la civilizaţia vecină este că nu avem indicii nici în privinţa probabilităţii ca viaţa inteligentă să apară într-un anumit loc. După cum a arătat astronomul american Frank Drake, această probabilitate poate fi calculată înmulţind probabilitatea ca acel mediu să fie locuibil (o planetă potrivită) cu probabilitatea ca acolo să apară viaţa şi cu probabilitatea ca viaţa să evolueze pentru a deveni inteligentă. Pe când eram student, n-aveam nici un indiciu privind vreuna dintre aceste trei probabilităţi. în ultimele două decenii au fost descoperite

planete orbitând în jurul altor stele, aşa încât acum pare probabil să existe o mulţime de planete locuibile, pesemne miliarde doar în galaxia noastră. Cât priveşte însă probabilitatea de apariţie a vieţii, iar apoi a inteligenţei, persistă o mare incertitudine: unii experţi cred că ambele sunt aproape inevitabile şi apar pe majoritatea planetelor locuibile, în vreme ce alţii cred că una din ele sau ambele sunt extrem de rare din cauza uneia sau mai multor „gâtuiri" evolutive care presupun un noroc extraordinar pentru a fi depăşite. Pentru unii, „gâtuirile" implică dificultăţi de tipul „oul sau găina" în etapele iniţiale ale vieţii care se autoreproduce: de pildă, pentru ca o celulă modernă să construiască un ribozom, mecanismul molecular complex care citeşte codul nostru genetic şi construieşte proteinele, este nevoie de un alt ribozom, şi nu e evident că primul ribozom poate evolua treptat de la ceva mai simplu.[10] Alţii cred că „gâtuirile" aduc cu sine dezvoltarea inteligenţei superioare. De exemplu, deşi dinozaurii au dominat Pământul vreme de peste 100 de milioane de ani, o perioadă de o mie de ori mai lungă decât timpul scurs de la apariţia oamenilor, evoluţia nu pare să-i fi împins spre o inteligenţă superioară şi inventarea telescoapelor sau calculatoarelor.

Unii resping argumentul meu susţinând că, intr-adevăr, viaţa inteligentă *ar putea* fi foarte rară, dar de fapt nu este – galaxia noastră e plină de viaţa inteligentă pe care oamenii de ştiinţă pur şi simplu n-o observă. Poate că extratereştrii au vizitat deja Pământul, după cum afirmă pasionaţii de OZN-uri. Poate că extratereştrii n-au vizitat Pământul, dar ei există undeva şi în mod deliberat se ascund de noi (aceasta a fost numită „ipoteza zoo" de astronomul american John A. Ball, şi apare în cărţi SF clasice cum e *StarMakerde* Olaf Stapledon). Sau poate că se află acolo, dar nu se ascund în mod deliberat: pur şi simplu nu-i interesează colonizarea spaţiului sau proiectele inginereşti la scară mare pe care le-am fi putut observa.

Sigur, nu trebuie să excludem aceste posibilităţi, însă din moment ce nu există dovezi general acceptate pentru vreuna dintre ele, trebuie să luăm în serios şi alternativa: suntem singuri. Apoi, cred că nu trebuie subestimată diversitatea civilizaţiilor avansate presupunând că toate împărtăşesc obiective care le fac să treacă neobservate: am văzut mai sus că obţinerea resurselor este un obiectiv cât se poate de firesc pentru o civilizaţie, iar noi am observa chiar şi dacă o *singură* civilizaţie s-ar hotărî să colonizeze tot ce-i stă la dispoziţie, răspândindu-se astfel în galaxie. Având în vedere că în galaxia noastră există milioane de planete locuibile asemănătoare Pământului, cu miliarde de ani mai bătrâne decât acesta, ai căror locuitori au avut suficient timp pentru a coloniza galaxia, nu putem respinge interpretarea cea mai naturală: apariţia vieţii este atât de improbabilă, încât toate acele planete sunt nelocuite.

Am putea afla curând dacă viaţa *nu* e un fenomen rar. Astronomii studiază atmosferele planetelor asemănătoare Pământului, încercând să obţină dovezi ale prezenţei oxigenului produs de viaţă. în paralel cu această căutare a *oricăror forme* de viaţă, căutarea vieţii *inteligente* a fost recent intensificată prin proiectul „Breakthrough Listen" finanţat cu roo de milioane de dolari de filantropul rus Iuri Milner.[§§§§§§§§§§]

Nu trebuie să fim prea antropocentrici atunci când căutăm forme avansate de viaţă: dacă descoperim o civilizaţie extraterestră, este probabil ca ea să fi devenit deja superinteligentă. într-un recent eseu, Martin Rees spunea: „Istoria civilizaţiei tehnologice umane se măsoară în secole – şi s-ar putea ca peste unul sau două secole oamenii să fie depăşiţi de inteligenţa anorganică, care va

§§§§§§§§§§ Lansat în 2015, proiectul vizează „ascultarea" cu ajutorul radiotele- scoapelor a unor eventuale semnale de telecomunicaţie emise de forme de viaţă extraterestre inteligente. (JV. t)

continua să evolueze vreme de miliarde de ani. [...] Este extrem de improbabil ca noi s-o «prindem» în scurtul interval de timp în care a luat formă organică."[11] Sunt de acord cu concluzia lui Jay Olson din articolul despre colonizarea spaţiului menţionat mai sus: „Considerăm că posibilitatea ca o inteligenţă avansată să folosească resursele universului doar pentru a popula planete asemănătoare Pământului cu versiuni perfecţionate ale oamenilor este un deznodământ improbabil al progresului tehnologic." Aşadar, dacă vă imaginaţi extratereştri, nu vă gândiţi la omuleţi verzi cu două mâini şi două picioare, ci la formele de viaţă superinte- ligente călătorind prin spaţiu, despre care am vorbit la începutul acestui capitol.

Deşi susţin cu putere toate încercările de a găsi viaţă extraterestră, care caută să răspundă Ia cele mai fascinante întrebări din ştiinţă, sper în taină ca ele să nu conducă la nimic! Aparenta incompatibilitate între abundenţa planetelor locuibile din galaxia noastră şi absenţa vizitatorilor extratereştri, cunoscută şi ca *paradoxul Iui Fermi,* sugerează existenţa a ceea ce economistul Robin Hanson numeşte „marele filtru", un baraj evolutiv/tehnologic undeva de-a lungul căii de dezvoltare de la materia nevie la viaţa care colonizează spaţiul. Dacă am descoperi în altă parte viaţă care a evoluat independent, asta ar însemna pesemne că viaţa primitivă nu e rară şi că barajul apare după stadiul nostru actual de dezvoltare – fie pentru că e imposibilă colonizarea spaţiului, fie pentru că aproape toate civilizaţiile avansate se autodistrug înainte de a putea ieşi în cosmos. Sper aşadar ca toate încercările de a găsi viaţă extraterestră să eşueze: aceasta ar fi în acord cu scenariul în care evoluţia vieţii inteligente este rară, dar noi, oamenii, am avut noroc, aşa încât am depăşit barajul şi avem un potenţial viitor extraordinar.

PERSPECTIVE

Am vorbit până acum despre istoria vieţii în universul nostru, de la începuturile ei modeste din urmă cu miliarde de ani la posibilul ei viitor măreţ de peste miliarde de ani. Dacă actuala dezvoltare a inteligenţei artificiale va declanşa o explozie a inteligenţei şi va optimiza colonizarea spaţiului, va fi cu adevărat o explozie cosmică: după ce a petrecut miliarde de ani ca o perturbaţie minusculă, aproape neglijabilă, într-un cosmos inert, viaţa explodează brusc pe arena cosmosului ca o undă de şoc sferică extinzându-se cu o viteză apropiată de cea a luminii, fără să încetinească vreodată, aprinzând totul în calea ei cu scânteia vieţii.

Asemenea idei optimiste privind importanţa vieţii din viitorul nostru cosmic au fost exprimate elocvent de mulţi dintre gânditorii pe care i-am întâlnit in această carte. Paradoxal, ţinând cont de posibilităţile superinteligenţei, majoritatea previziunilor autorilor de SF, adesea consideraţi visători romantici nerealişti, şi ale oamenilor de ştiinţă privind colonizarea spaţiului par acum prea *pesimiste.* De pildă, am văzut că devine mult mai uşoară călătoria intergalactică odată ce fiinţele umane şi alte entităţi inteligente se pot propaga sub formă digitală, ceea ce ne dă prilejul să devenim stăpâni pe propriul nostru destin nu numai în sistemul solar sau în Calea Lactee, dar şi în cosmos.

Am luat în considerare mai sus posibilitatea absolut reală de a fi singura civilizaţie cu tehnologie avansată din universul nostru, în încheierea acestui capitol, vă propun să discutăm despre acest scenariu şi despre uriaşa responsabilitate morală pe care o implică. După 13,8 miliarde de ani de la big bang, viaţa din universul nos tru a ajuns la o bifurcaţie: are în faţă fie înflorirea la scară cosmică, fie extincţia. Dacă nu ne perfecţionăm tehnologia, problema care se pune nu este dacă omenirea va dispărea, ci *cum.* Ce ne va lovi mai întâi – un asteroid, erupţia unui

supervulcan, căldura arzătoare a Soarelui care îmbătrâneşte sau vreo altă calamitate (vezi figura 5.1)? Odată ce vom dispărea, drama cosmică prezisă de Freeman Dyson se va juca fără spectatori: stelele îşi vor epuiza combustibilul, galaxiile vor păli, iar găurile negre se vor evapora, fiecare sfârşindu-şi viaţa cu o explozie uriaşă care va elibera de peste un milion de ori mai multă energie decât „Bomba Ţarului“, cea mai puternică bombă cu hidrogen construită vreodată. „Expansiunea rece a universului va fi pentru foarte multă vreme luminată din când în când de focuri de artificii”, spune Freeman. Nu va fi însă nimeni care să admire spectacolul acestor focuri de artificii.

Fără tehnologie, extincţia omenirii e iminentă la scara cosmică a miliardelor de ani, făcând ca întreaga dramă a vieţii în universul nostru să fie doar o strălucire efemeră de frumuseţe, pasiune şi sens într-o cvasietemitate de nonsens trăită de nimeni. Ce ocazie irosită ar fi! Dacă în loc să respingem tehnologia am alege s-o îmbrăţişăm, atunci am mări miza jocului: ar creşte şansele şi ca viaţa să dăinuie şi să înflorească, şi ca viaţa să dispară încă mai repede, autodistrugându-$e din cauza unei slabe capacităţi de a prevedea (vezi figura 5.1). Eu votez pentru îmbrăţişarea tehnologiei şi continuarea drumului, călăuziţi de luciditate, precauţie şi planificare atentă.

După 13,8 miliarde de ani de istorie cosmică, ne aflăm într-un univers de o neasemuită frumuseţe, care prin noi, oamenii, a devenit viu şi a căpătat conştiinţă de sine. Am văzut că potenţialul viitor al vieţii în universul nostru e mult mai măreţ decât ar fi putut visa strămoşii noştri, dar există în acelaşi timp riscul ca viaţa inteligentă să dispară pentru totdeauna. Se va împlini oare potenţialul vieţii din universul nostru, sau va fi el irosit? Asta depinde în mare măsură de ceea ce facem noi, oamenii care trăim azi, în cursul vieţii noastre, iar eu cred că putem face ca viitorul vieţii să fie într-adevăr extraordinar, cu condiţia să luăm

deciziile corecte. Ce-ar trebui să ne dorim şi cum putem atinge acele obiective? Vă propun ca în restul cărţii să examinăm unele dintre cele mai dificile provocări cu care ne confruntăm şi căile de urmat.

REZUMAT:

• La scara de miliarde de ani a timpului cosmic, o explozie a inteligenţei este un eveniment brusc în care tehnologia ajunge rapid la limitele impuse doar de legile fizicii.

• Tehnologia aflată la această limită e mult mai înaltă decât tehnologia de azi, permiţând ca o cantitate dată de materie să genereze de vreo zece miliarde de ori mai multă energie (folosind sfaleroni sau găuri negre), să stocheze cu 12-18 ordine de mărime mai multă informaţie sau să calculeze cu 31-41 ordine de mărime mai repede – sau să fie transformată în orice altă formă de materie se doreşte.

• Viaţa superinteligentă nu numai că ar folosi incomparabil mai eficient resursele existente, dar ar putea face şi ca biosfera de azi să crească cu aproximativ 32 de ordine de mărime, dobândind noi resurse prin colonizarea cosmică la o viteză apropiată de cea a luminii.

• Energia întunecată limitează expansiunea cosmică a vieţii superinteligente, dar o şi protejează de bulele mortale şi de civilizaţiile ostile aflate la mare distanţă. Ameninţarea energiei întunecate care separă civilizaţiile cosmice impune conceperea unor ample proiecte de inginerie cosmică, între care construirea unor găuri de vierme, presupunând că aceasta e cu putinţă.

• Informaţia este probabil să devină principalul bun care să fie partajat sau vândut la distanţe cosmice.

• Făcând abstracţie de găurile de vierme, limita impusă de viteza luminii asupra comunicaţiei îngreunează coordonarea şi controlul în cadrul unei civilizaţii cosmice. Un centru aflat la mare distanţă îşi poate determina „nodurile“ superinteligente să coopereze fie prin recompense, fie prin ameninţări – de pildă, cu un gardian IA programat să distrugă nodul declanşând o supernovă sau un quasar dacă regulile sunt încălcate.

■ Confruntarea a două civilizaţii în expansiune poate duce la asimilare, cooperare sau război, acesta din urmă fiind pesemne mai puţin probabil decât este între civilizaţiile de azi.

• în ciuda credinţei contrare extrem de răspândite, este perfect plauzibil să fim singura formă de viaţă capabilă să însufleţească în viitor universul nostru observabil.

• Dacă nu ne perfecţionăm tehnologia, problema care se pune nu este dacă omenirea va dispărea, ci cum: ce ne va lovi mai întâi, un asteroid, erupţia unui supervulcan, căldura arzătoare a Soarelui care îmbătrâneşte sau vreo altă calamitate?

• Dacă ne vom perfecţiona în continuare tehnologia cu suficientă precauţie pentru a evita capcanele, viaţa are posibilitatea să înflorească pe Pământ şi mult dincolo de el vreme de miliarde de ani, aşa cum strămoşii noştri nici măcar nu puteau visa.

CAPITOLUL/

Obiective

> *Taina existenţei umane nu constă în a trăi, ci în a şti pentru ce trăieşti.* — Feodor Dostoievsky *Fraţii Karamazov*
>
> *Viaţa e o călătorie, nu o destinaţie.* — Ralph Waldo Emerson

Dacă ar fi să rezum cu un singur cuvânt subiectul celor mai aprige controverse legate de IA, acesta ar fi „obiective“. Ar trebui oare să înzestrăm LA cu obiective, şi care ar fi acele obiective? Cum putem înzestra IA cu obiective? Cum putem fi siguri că aceste obiective sunt păstrate, chiar dacă IA devine mai deşteaptă? Putem schimba obiectivele unei IA care e mai deşteaptă decât noi? Care sunt obiectivele noastre supreme? Aceste întrebări sunt nu numai dificile, dar ele au o importanţă crucială pentru viitorul vieţii: dacă nu ştim ce vrem, e mai puţin probabil să obţinem acel lucru, iar dacă cedăm controlul unor maşini care nu ne împărtăşesc obiectivele, vom obţine probabil ceea ce nu vrem.

FIZICĂ: ORIGINEA OBIECTIVELOR

Pentru a lămuri aceste probleme, să examinăm mai întâi originea ultimă a obiectivelor. Când privim în jurul nostru,

unele procese par *orientate spre obiective,* altele nu. De exemplu, procesul în care o minge de fotbal e lovită cu piciorul pentru a se înscrie golul decisiv. Comportamentul mingii în sine nu pare orientat spre vreun obiectiv, şi se explică cel mai simplu pe baza legilor newtoniene ale mişcării,
ca reacţie la lovitura aplicată de picior. Pe de altă parte, comportamentul fotbalistului se explică cel mai simplu nu din perspectivă mecanică, prin atomi care se izbesc unii de alţii, ci prin faptul că fotbalistul are *obiectivul* dea înscrie pentru echipa lui. Cum a apărut un asemenea comportament orientat spre obiective din fizica universului nostru timpuriu, care se reducea Ia doar câteva tipuri de particule ciocnindu-se unele de altele fără vreun obiectiv aparent?

în mod surprinzător, rădăcinile ultime ale comportamentului orientat spre obiective pot fi găsite în chiar legile fizicii şi se manifestă până şi în cel mai simplu proces care nu implică viaţa. Dacă un salvamar porneşte în salvarea unui înotător, cum se vede în figura 7.1, ne aşteptăm ca acesta să nu pornească în linie dreaptă, ci să alerge puţin de-a lungul plajei, unde se deplasează mai repede decât în apă, iar apoi să se întoarcă puţin pentru a se arunca in mare. Considerăm în chip firesc că această alegere a traiectoriei este orientată spre un obiectiv, fiindcă dintre toate traiectoriile posibile, salvamarul o alege în mod deliberat pe cea optimă pentru a ajunge cât mai repede la înotător. Şi totuşi, o simplă rază de lumină se „îndoaie“ în mod asemănător atunci când intră în apă (vezi figura 7.1), minimizând şi ea timpul de călătorie până la destinaţie. Cum e posibil aşa ceva?

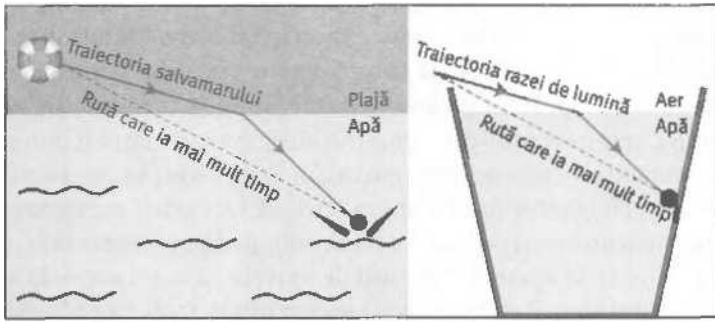


Figura 7.1 Pentru a salva un înotător cât mai rapid cu putinţa, un salvamar nu se va mişca în linie dreaptă (linia punctată), ci va alerga puţin de-a lungul plajei, unde se deplasează mai repede decât în apă. 0 rază de lumină se „îndoaie"în mod asemănător la intrarea în apă, pentru a ajunge la destinaţie cât mai rapid cu putinţă.

în fizică, acesta se numeşte *principiul lui Fermat,* formulat în 1662, şi oferă o modalitate alternativă de a prezice comportamentul razelor de lumină. De atunci, fizicienii au descoperit că *toate* legile fizicii clasice pot fi formulate matematic într-un mod asemănător: dintre toate căile pe care natura le-ar putea alege pentru a face ceva, ea preferă calea optimă, care în general se reduce la minimizarea sau maximizarea unei anumite cantităţi. Există două modalităţi echivalente matematic de a prezenta orice lege din fizică: fie că trecutul determină viitorul, fie că natura optimizează ceva. Deşi a doua modalitate nu e de regulă predată la cursurile introductive pentru că presupune dificultăţi matematice, mie mi se pare mai elegantă şi mai profund! Dacă cineva încearcă să optimizeze ceva (de pildă, scorul, averea sau fericirea), vom spune că demersul lui e orientat spre un obiectiv. Deci dacă natura însăşi încearcă să optimizeze ceva, nu-i de mirare că poate apărea un comportament orientat spre un obiectiv: acesta a fost încorporat de la bun început în înseşi legile fizicii.

O celebră mărime pe care natura se străduieşte s-o maximizeze este *entropia,* care, simplu spus, măsoară gradul de dezordine. A doua lege a termodinamicii afirmă

că entropia tinde să crească până îşi atinge valoarea maximă posibilă. Ignorând deocamdată efectele gravitaţiei, această stare finală de dezordine maximă se numeşte *moarte termică* – totul e împrăştiat într-o uniformitate anostă absolută, nu există complexitate, viaţă sau schimbare. Când tomi lapte rece în cafea fierbinte, de pildă, lichidul pare să se îndrepte ireversibil către propriul iui obiectiv al morţii termice, iar curând e un amestec uniform călduţ. Dacă un organism moare, entropia lui începe de asemenea să crească, iar curând aranjamentul particulelor sale tinde să devină mult mai puţin organizat.

Obiectivul aparent al naturii de a creşte entropia explică de ce timpul pare să aibă o direcţie preferenţială, făcând ca filmele să fie nerealiste dacă sunt derulate invers: dacă scapi un pahar cu vin, te aştepţi sa se izbească de podea şi să se spargă, dezordinea globală (entropia) crescând. Dacă vezi pe urmă că cioburile se adună şi recompun paharul care sare înapoi intact în mâna ta (scăzând entropia), probabil că nu l-ai bea şi ţi-ai spune că eşti beat.

Când am aflat prima oară despre înaintarea inexorabilă către moartea termică, am găsit că e un lucru deprimant, senzaţie împărtăşită şi de lordul Kelvin, unul dintre întemeietorii termodinamicii, care scria în 1841 că „rezultatul ar fi inevitabil o stare de repaus şi moarte universale“. E greu să te împaci cu ideea că obiectivul pe termen lung al naturii e moartea şi distrugerea. Descoperiri recente au arătat însă că lucrurile nu stau chiar atât de rău. în primul rând, gravitaţia se comportă diferit faţă de toate celelalte forţe, tinzând să facă universul nostru nu mai uniform şi mai anost, ci mai structurat şi mai interesant. Gravitaţia a transformat anostul univers timpuriu, care era aproape perfect uniform, în cosmosul de azi structurat şi minunat de complex, plin de galaxii, stele şi planete. Datorită gravitaţiei există acum o gamă largă de temperaturi, permiţând vieţii să înflorească prin

combinaţia dintre cald şi rece: trăim pe o planetă confortabilă, care absoarbe căldură de la Soarele aflat la temperatura de 6 ooo° C şi radiază surplusul de căldură în spaţiul rece, a cărui temperatură este de doar 3^0 C deasupra lui zero absolut.

în al doilea rând, cercetări recente ale colegului meu de la MIT Jeremy England şi ale altora arată că termodinamica oferă naturii şi un obiectiv mai interesant decât moartea termică.[1] Acest obiectiv se leagă de ceea ce numim *adaptare determinată de disipare (dissipation-driven adaptatiori):* un grup aleatoriu de particule tinde să se organizeze astfel încât să extragă din mediu energie cât mai eficient cu putinţă („disipare" înseamnă creşterea entropiei transformând de regulă energia utilă în căldură şi efectuând adesea lucru mecanic util în acest proces). De pildă, un grup de molecule expuse la lumina Soarelui are tendinţa ca, în timp, să se aranjeze astfel încât să absoarbă cât mai bine lumina. Cu alte cuvinte, natura pare să fie înzestrată cu obiectivul de a produce sisteme care se auto-organizează din ce în ce mai complexe, asemănătoare vieţii, iar acest obiectiv e încorporat în înseşi legile fizicii.

Cum putem reconcilia această tendinţă cosmică spre viaţă cu tendinţa cosmică spre moarte termică? Răspunsul se află în celebra carte din 1944 *Ce este viaţa?* a lui Erwin Schrodinger, unul dintre creatorii mecanicii cuantice. Schrodinger subliniază că semnul distinctiv al sistemelor vii este că ele menţin sau reduc entropia crescând entropia din jurul lor. Altfel spus, legea a doua a termodinamicii conţine un amendament favorabil vieţii: deşi entropia totală trebuie să crească, ea poate scădea în anumite locuri atâta timp cât creşte şi mai mult în altele. Viaţa îşi menţine sau îşi sporeşte aşadar complexitatea făcând ca mediul să devină şi mai dezordonat.

Am văzut că originea comportamentului orientat spre obiective poate fi găsită în legile fizicii, care par să ofere particulelor obiectivul de a se aranja astfel încât să extragă energie din mediu cât mai eficient cu putinţă. O soluţie bună pentru ca un aranjament de particule să-şi urmărească mai departe acest obiectiv este ca aranjamentul să facă copii ale Iui însuşi, să producă mai mulţi absorbanţi de energie. Se cunosc numeroase exemple de asemenea auto-replicare emergentă: vârtejuri din fluide turbulente fac copii ale lor însele, iar aglomerări de miscrosfere pot determina microsferele din vecinătate să formeze aglomerări identice. La un moment dat, un anumit aranjament de particule s-a perfecţionat atât de mult în copierea lui însuşi, încât a putut s-o facă aproape la nesfârşit extrăgând din mediul lui energie şi materii prime. Numim acest aranjament de particule *viaţă.* Ştim deocamdată foarte puţin despre felul în care a apărut viaţa pe Pământ, cert este că forme primitive de viaţă existau deja cu vreo 4 miliarde de ani în urmă.

Dacă o formă de viaţă se copiază pe ea însăşi, iar copiile fac acelaşi lucru, atunci numărul total se va dubla la intervale regulate, până când dimensiunea populaţiei se loveşte de limitările resurselor sau de alte dificultăţi. Dublarea repetată produce curând numere uriaşe: dacă începi cu unu şi dublezi de doar trei sute de ori, ajungi la ceva care depăşeşte numărul particulelor din universul nostru. Asta înseamnă că, la scurt timp după apariţia primei forme de viaţă primitivă, cantităţi imense de materie au căpătat viaţă. Uneori copierea nu era perfectă, aşa încât existau mai multe forme de viaţă care încercau să se copieze pe ele insele, concurând pentru aceleaşi resurse limitate. începuse evoluţia darwinistă.

Dacă ai fi putut observa Pământul în momentul apariţiei vieţii, ai fi remarcat o schimbare spectaculoasă în

comportamentul orientat spre obiective. Dacă înainte particulele păreau că încearcă să sporească pe diferite căi dezordinea medie, aceste noi structuri care se autocopiau păreau să aibă un alt obiectiv: nu disiparea, ci *replicarea.* Charles Darwin a dat o explicaţie elegantă: din moment ce majoritatea copiatorilor eficienţi îi surclasează şi îi domină pe ceilalţi, curând orice formă de viaţă pe care o vezi va fi extrem de optimizată pentru obiectivul replicării.

Cum s-a putut schimba obiectivul de la disipare la replicare, atunci când legile fizicii au rămas aceleaşi? Răspunsul este că obiectivul fundamental (disiparea) *nu* s-a schimbat, ci a condus la un alt *obiectiv instrumental* -un subobiectiv care a ajutat la îndeplinirea obiectivului fundamental. Să luăm ca exemplu mâncatul. Părem cu toţii să avem obiectivul de a ne potoli foamea, dar ştim că singurul obiectiv fundamental al evoluţiei e replicarea, nu masticarea. Aceasta pentru că mâncatul ajută replicarea: dacă mori prin înfometare nu poţi avea copii. în mod asemănător, replicarea ajută disiparea, fiindcă o planetă mustind de viaţă disipă energia mai eficient. într-un anume sens deci, cosmosul nostru a inventat viaţa pentru a-l ajuta să se apropie mai repede de moartea termică. Dacă verşi zahăr pe podeaua bucătăriei, acesta îşi poate în principiu păstra energia chimică utilă ani de-a rândul, dar dacă apar furnicile, ele vor disipa imediat acea energie. La fel, rezervele de petrol îngropate în crusta Pământului şi-ar păstra energia lor chimică utilă mult mai mult timp dacă noi, forme de viaţă bipede, nu l-am extrage şi nu l-am arde.

Pentru locuitorii evoluaţi ai Pământului din prezent, aceste obiective instrumentale par să fi căpătat propria lor viaţă: deşi evoluţia le-a optimizat pentru unicul obiectiv al replicării, mulţi îşi petrec o mare parte din timp nu producând urmaşi, ci în activităţi cum sunt somnul, căutarea hranei, construirea de locuinţe, impunerea dominaţiei şi lupta cu semenii sau ajutorarea lor- uneori chiar într-o măsură care *reduce* replicarea. Cercetările din

psihologia evoluţionistă, ştiinţele economice şi inteligenţa artificială explică în mod elegant de ce. Unii economişti au construit modele în care oamenii acţionează ca agenţi raţionali, idealizând luarea deciziilor prin presupunerea că oamenii aleg întotdeauna ce este optim pentru urmărirea obiectivului lor. Evident, asemenea modele sunt nerealiste. În practică, aceşti agenţi au ceea ce Herbert Simon, laureat Nobel şi figură marcantă a IA, numea „raţionalitate mărginită“, fiindcă posedă resurse limitate: raţionalitatea lor e limitată de informaţia disponibilă, timpul pe care îl au la dispoziţie pentru a se gândi şi hardware-ul disponibil cu care gândesc. Asta înseamnă că atunci când evoluţia darwinistă optimizează un organism pentru a atinge un obiectiv, în cel mai bun caz ceea ce poate face este să pună în aplicare un algoritm aproximativ care funcţionează acceptabil în contextul restrâns în care agentul se găseşte de regulă. Evoluţia a pus în aplicare optimizarea replicării exact în acest mod: în loc să judece în fiecare situaţie ce acţiune va maximiza numărul de progenituri reuşite ale organismului, foloseşte un amestec eterogen de reguli empirice care de obicei funcţionează bine. Pentru majoritatea animalelor, printre acestea se numără a avea dorinţe sexuale, a bea atunci când ţi-e sete, a mânca atunci când ţi-e foame şi a evita lucrurile care au un gust rău sau te rănesc.

Aceste reguli empirice dau uneori greş în situaţii pentru care n-au fost concepute, ca atunci când şobolanii mănâncă o otravă cu gust delicios, moliile sunt atrase pe o bandă lipicioasă de mirosul seducător de femelă, iar cărăbuşii sunt arşi de flacăra unei lumânări.[**********] Din moment ce societatea umană de azi e foarte diferită de

---

[**********] O regulă empirică folosită de numeroase insecte pentru a zbura în linie dreaptă este de a presupune că lumina vine de la Soare şi de a zbura sub un unghi fix în raport cu ea. Dacă lumina provine de Ia o flacără din apropiere, cărăbuşul e din nefericire păcălit şi se îndreaptă în spirală către moarte. (N. a.)

mediul pentru care evoluţia a optimizat regulile noastre empirice, n-ar trebui să ne mire că adesea comportamentul nostru nu reuşeşte să maximizeze numărul progeniturilor. De pildă, subobiectivul nostru de a nu muri de foame este parţial pus în aplicare ca dorinţă de a consuma o hrană bogată în calorii, declanşând epidemia actuală de obezitate şi dificultăţi în curtare. Subobiectivul de a procrea a fost pus în aplicare ca dorinţă sexuală, nu ca dorinţă de a deveni donator de spermă sau de ovule, deşi aceasta poate produce mai mulţi copii cu efort mai mic.

PSIHOLOGIE: URMĂRIREA OBIECTIVELOR ŞI REVOLTA ÎMPOTRIVA LOR

în concluzie, un organism viu este un agent cu raţionalitate mărginită care nu urmăreşte un singur obiectiv, ci aplică reguli empirice pentru ce trebuie să urmărească şi ce trebuie să evite. Minţile noastre umane percep aceste reguli empirice evoluate ca pe *senzaţii* şi *sentimente,* care de regulă (iar adesea fără să fim conştienţi) călăuzesc deciziile noastre către obiectivul suprem al replicării. Senzaţiile de foame şi sete ne apără de moartea prin înfometare şi deshidratare, senzaţiile de durere ne protejează corpurile, senzaţiile voluptuoase ne fac să procreăm, sentimentele de iubire şi compasiune ne fac să-i ajutăm pe alţi purtători ai genelor noastre sau pe cei care îi ajută pe aceştia etc. Călăuzit de aceste senzaţii şi sentimente, creierul nostru poate hotărî rapid şi eficient ce să facă, fără a supune fiecare alegere unei analize îndelungate a consecinţelor ultime asupra numărului de descendenţi pe care îi vom produce. Pentru perspective asemănătoare privind senzaţiile, sentimentele şi rădăcinile lor psihologice, recomand călduros lucrările lui William James şi Antonio Damasio.[2]

E important de observat că atunci când senzaţiile şi sentimentele noastre *se opun* procreării, nu e vorba

neapărat de un accident sau de o înşelătorie: creierul nostru se poate revolta în mod deliberat împotriva genelor noastre şi a obiectivului lor de replicare, folosind, de pilda, mijloace contraceptive! Exemple extreme de revoltă a creierului împotriva genelor sunt sinuciderea sau alegerea de a-ţi petrece viaţa în celibat ca preot, călugăr sau călugăriţă.

De ce alegem uneori să ne revoltăm împotriva genelor şi a obiectivului lor de replicare? Ne revoltăm fiindcă, prin însăşi firea noastră, ca agenţi cu raţionalitate mărginită, suntem loiali numai sentimentelor şi senzaţiilor noastre. Deşi creierul a evoluat numai ca să ajute la copierea genelor noastre, nu-i pasă de acest obiectiv, din moment ce nu simte nimic pentru gene – de fapt, în cea mai mare parte a istoriei umane, strămoşii noştri nici măcar nu ştiau că au gene. Apoi, creierul e mult mai deştept decât genele noastre, iar acum că înţelegem obiectivul genelor (replicarea), îl găsim banal şi lesne de ignorat. Oamenii îşi pot da seama de ce genele îi îndeamnă spre sex, şi totuşi nu-şi doresc să crească cincisprezece copii, aşa încât aleg să-şi deturneze programarea genetică şi combină răsplata afectivă a intimităţii cu controlul natalităţii. Ei pot înţelege de ce genele îi fac să saliveze după dulciuri, însă nu vor să se îngraşe, iar astfel aleg să-şi deturneze programarea genetică şi combină răsplata senzorială de a mânca dulciuri cu îndulcitorii artificiali fără calorii.

Deşi asemenea deturnări ale mecanismului recompenselor dau uneori greş, ca atunci când oamenii devin dependenţă de heroină, bazinul genelor umane a supravieţuit destul de bine, în ciuda cre- ierelor noastre viclene şi rebele. E important să ţinem totuşi minte că acum autoritatea supremă sunt sentimentele şi senzaţiile, nu genele noastre. Comportamentul uman nu este aşadar strict optimizat pentru supravieţuirea speciei noastre. De fapt, din moment ce sentimentele şi senzaţiile noastre pun

în aplicare doar reguli empirice care nu sunt adaptate tuturor situaţiilor, comportamentul uman, strict vorbind, nu are un unic obiectiv bine definit.

INGINERIE: EXTERNAL1ZAREA OBIECTIVELOR

Pot avea maşinile obiective ? Această întrebare simplă a declanşat ample dezbateri pentru că i s-au dat semnificaţii diferite, adesea legate de probleme spinoase – pot fi maşinile conştiente, pot ele avea sentimente? Dacă suntem însă mai pragmatici şi ne întrebăm
pur şi simplu dacă pot manifesta maşinile un comportament orientat spre obiective, atunci răspunsul e evident: „Desigur că pot, din moment ce Ie putem proiecta astfel!“ Proiectăm curse de şoa reci cu obiectivul de a prinde şoareci, maşini de spălat vasele cu obiectivul de a spăla vasele şi ceasuri cu obiectivul de a măsura timpul. Când ai de-a face cu o maşină, faptul empiric că manifestă un comportament orientat spre un obiectiv este de regulă singurul lucru care contează: dacă eşti urmărit de o rachetă termoghidată, nu-ţi pasă dacă are conştiinţă sau sentimente! în cazul în care to tuşi nu-ţi convine să spui că racheta are un obiectiv, deşi nu e conştientă, poţi citi de-acum înainte „scop“ atunci când eu scriu „obiectiv“ – despre conştiinţă vom vorbi în capitolul următor.

Majoritatea lucrurilor pe care le-am construit până acum au fost doar *proiectate în* vederea unui obiectiv, nu manifestă un *comportament* orientat spre un obiectiv: o autostradă n-are vreun comportament, ea se află pur şi simplu acolo. Cea mai economică explicaţie pentru existenţa ei este însă că a fost proiectată pentru a îndeplini un obiectiv, prin urmare chiar şi asemenea tehnologie pasivă face ca universul nostru să fie într-o măsură mai mare orientat spre obiective. *Tehnologia* este explicarea lucrurilor prin scopurile lor, nu prin cauzele lor, aşa încât

putem rezuma prima parte a acestui capitol spunând că universul nostru devine tot mai tehnologic.

Nu numai că materia neînsufleţită poate avea obiective, cel puţin în acest sens slab, dar chiar asta se întâmplă, în măsură tot mai mare. Dacă ai fi observat atomii Pământului încă de Ia formarea planetei, ai fi observat trei etape de comportament orientat spre obiective:

i. întreaga materie părea concentrată asupra disipării (creşterea entropiei).
2. 0 parte din materie a căpătat viaţă şi s-a concentrat în schimb asupra replicării şi subobiectivelor ei.
3. 0 parte rapid crescătoare a materiei a fost rearanjată de organismele vii pentru a ajuta la îndeplinirea obiectivelor lor.

Tabelul 7.1 arată cât de dominantă a devenit omenirea din perspectiva fizicii: nu numai că am ajuns să conţinem mai multă materie decât toate celelalte mamifere, cu excepţia vacilor (care sunt atât de numeroase fiindcă slujesc obiectivelor noastre de a consuma came de vită şi produse lactate), dar materia din maşinile, drumurile, clădirile noastre şi din alte proiecte inginereşti pare să se apropie de întreaga materie vie de pe Pământ. Cu alte cuvinte, chiar şi fără o explozie a inteligenţei, majoritatea materiei de pe Pământ care prezintă proprietăţi orientate spre obiective va fi în curând rezultatul proiectării, nu al evoluţiei.

Entităţi orientate spre obiective	**Miliarde de tone**
5 x io3° bacterii	400
Plante	400
10^ peşti mezopelagici	10
1,3 x io^9 vaci	0,5
7 x 10^9 oameni	0,4
10^{14} furnici	0,3

1,7 x io^6 balene	0,0005
Beton	100
Oţel	20
Asfalt	15
1,2 x 10^9 automobile	2

Tabelul 7.1 Cantităţile aproximative de materie de pe Pământ din entităţi care au evoluat sau au fost proiectate pentru un obiectiv. Entităţile făurite de mâna omului, cum ar fi clădiri, drumuri şi automobile, par să se apropie cantitativ de entităţile care au evoluat, cum sunt plantele şi animalele.

Acest recent al treilea tip de comportament orientat spre obiective poate deveni mult mai variat decât precedentul: dacă toate entităţile care au evoluat au acelaşi obiectiv suprem (replicarea), entităţile proiectate pot avea practic orice obiective supreme, intrând chiar în contradicţie unele cu altele. Aragazele încearcă să încălzească hrana, în vreme ce frigiderele încearcă s-o răcească. Generatoarele încearcă să convertească mişcarea în electricitate, în vreme ce motoarele încearcă să convertească electricitatea în mişcare. Programele standard pentru jocul de şah încearcă să câştige partidele, însă există şi programe care participă la tumee în care obiectivul e să pierzi.

Tendinţa istorică este ca entităţile proiectate să aibă obiective nu doar mai diverse, dar şi mai *complexe:* dispozitivele noastre devin tot mai isteţe. Am construit primele maşini şi alte artefacte în virtutea unor obiective simple – de pildă, case care să ne ferească de frig, umezeală şi primejdii. Am învăţat treptat să construim maşini cu obiective mai complexe, cum sunt aspiratoarele de praf robotice, rachetele autonome, automobilele care se conduc singure. Progresul recent al IA ne-a oferit sisteme precum Deep Blue, Watson şi AlphaGo, ale căror obiective de a câştiga Ia şah, concursuri cu întrebări şi go sunt atât de complicate, încât e nevoie de cunoştinţe umane

remarcabile pentru a înţelege cu adevărat cât de capabile sunt aceste sisteme.

Când construim o maşină care să ne ajute, obiectivele ei pot fi greu de aliniat cu ale noastre. De pildă, o cursă de şoareci poate confunda degetele tale de la picioare cu un rozător înfometat, iar urmările sunt dureroase. Toate maşinile sunt agenţi cu raţionalitate mărginită, până şi cele mai sofisticate maşini de azi sunt departe de a înţelege lumea la fel ca noi, aşa încât regulile pe care le folosesc pentru a şti ce au de făcut sunt adesea extrem de simple. Acea cursă de şoareci se declanşează cu prea mare uşurinţă fiindcă n-are nici un indiciu care să-i spună ce e un şoarece, multe accidente industriale letale se produc fiindcă maşinile n-au nici un indiciu care să le spună ce este un om, iar calculatoarele care au declanşat „crahul fulger“ de un bilion de dolari din 2010 n-aveau nici un indiciu care să Ie spună că ceea ce făceau era absurd. Numeroase asemenea probleme de aliniere a obiectivelor pot fi deci rezolvate sporind inteligenţa maşinilor, însă aşa cum ne-a arătat Prometeu în capitolul 4, nici măcar o inteligenţă superioară a maşinilor nu ne poate asigura că acestea împărtăşesc obiectivele noastre.

IA PRIETENOASĂ: ALINIEREA OBIECTIVELOR

Cu cât maşinile devin mai inteligente şi mai puternice, cu atât e mai important ca obiectivele lor să fie aliniate cu ale noastre. Câtă vreme construim doar maşini relativ stupide, întrebarea nu e dacă obiectivele umane vor predomina în cele din urmă, ci ce neajunsuri vor provoca omenirii aceste maşini până când vom găsi o soluţie la problema alinierii obiectivelor. Dacă însă o superinte- ligenţă va ieşi vreodată de sub controlul nostru, lucrurile vor sta invers: din moment ce inteligenţa este capacitatea de a îndeplini obiective, prin definiţie o IA superinteligentă îşi va îndeplini mult mai bine obiectivele decât le îndeplinim noi pe ale noastre, devenind aşadar dominantă. Am analizat

numeroase asemenea exemple legate de Prometeu în capitolul 4. Dacă vrei să vezi cum te surclasează chiar azi obiectivele unei maşini, descarcă un program de ultimă generaţie pentru jocul de şah şi încearcă să-l învingi. N-ai să reuşeşti, iar programele se perfecţionează rapid...

Cu alte cuvinte, *adevăratul risc legat de IAG nu e reaua voinţă, ci competenţa. O* IA superinteligentă îşi va îndeplini obiectivele extrem de bine, iar dacă obiectivele ei nu sunt aliniate cu ale noastre, ne aflăm într-o situaţie complicată. După cum am arătat în capitolul 1, oamenii nu ezită să inunde muşuroaie de furnici pentru a construi baraje de hidrocentrale, prin urmare ar trebui ca omenirea să nu ajungă în situaţia acelor furnici. Majoritatea cercetătorilor susţin aşadar că, dacă vom ajunge vreodată să creăm o superinteligenţă, va trebui să ne asigurăm că este, folosind expresia lui Eliezer Yudkowsky, o „IA prietenoasă": obiectivele ei sunt aliniate cu ale noastre.[3]

Alinierea obiectivelor unei IA superinteligente cu cele ale noastre e un lucru nu doar important, dar şi dificil. De fapt, deocamdată rămâne o problemă nerezolvată. Ea se împarte în trei subprobleme, fiecare fiind un subiect de cercetare intensă pentru informaticieni şi specialişti din alte domenii:

1. Să determini IA să *înţeleagă* obiectivele noastre
2. Să determini IA să *adopte* obiectivele noastre
3. Să determini IA să *păstreze* obiectivele noastre

Să le analizăm pe rând, iar în secţiunea următoare vom lămuri ce ar putea însemna „obiectivele noastre".

Pentru a înţelege obiectivele noastre, o IA trebuie să priceapă nu ce facem, ci de ce facem ceea ce facem. Noi, oamenii, îndeplinim sarcina asta fără efort, aşa încât nu ne dăm seama cât de dificilă e pentru un calculator şi ce uşor se creează confuzii. Dacă ceri unui automobil din viitor care se conduce singur să te ducă la aeroport cât mai repede cu putinţă, iar automobilul ia ce spui *ad litteram,* vei ajunge acolo urmărit de helicoptere şi acoperit de

vomă. Dacă strigi „Nu asta am vrut!“, îţi poate răspunde pe drept cuvânt „Dar asta mi-ai cerut“. Acelaşi lucru se întâmplă în multe poveşti celebre. într-o legendă grecească din Antichitate, regele Midas cere ca tot ce atinge să se transforme în aur, dar ajunge să nu mai poată mânca, ba chiar îşi transformă fiica într-o statuie de aur. în poveştile în care un duh îndeplineşte trei dorinţe, există multe variante pentru primele două, dar aproape de fiecare dată a treia este: „Anulează, te rog, primele două dorinţe, pentru că nu asta voiam de fapt.“

Toate aceste exemple arată că, pentru a-ţi da seama ce vor oamenii în realitate, nu ajunge să te bazezi pe ce spun. Ai nevoie şi de un model detaliat al lumii, incluzând numeroasele preferinţe comune pe care de regulă nu le mai explicităm, fiindcă ni se par evidente – de pildă, nu vrem să vomităm sau să mâncăm aur. Odată ce avem un asemenea model al lumii, ne putem da seama adesea ce vor oamenii, chiar dacă nu ne-o spun, observând pur şi simplu comportamentul lor orientat spre obiective. Copiii ipo- criţilor învaţă mai multe din ce fac decât din ce spun părinţii lor.

Cercetătorii IA se străduiesc în prezent să doteze maşinile cu capacitatea de a deduce obiectivele din comportament, ceea ce va fi util şi cu mult înainte de apariţia unei superinteligenţe. De pildă, ar fi bine pentru o persoană în vârstă ca robotul care o îngrijeşte să-şi poată da seama ce-i place pur şi simplu observând-o, fără să fie nevoie să i se explice totul prin cuvinte sau prin pro- gramarea calculatorului. O problemă este codificarea într-un calculator a unor sisteme arbitrare de obiective şi principii etice, iar alta este construirea unor maşini care să-şi poată da seama ce sistem se potriveşte cel mai bine comportamentului observat.

O abordare a celei de-a doua probleme folosită în prezent este *învăţarea prin consolidare inversă (inverse reinforcement leaming),* principala preocupare a noului

centru de cercetări de la Berkeley înfiinţat de Stuart Russell. Să presupunem, de exemplu, că o IA vede cum un pompier intră într-o clădire în flăcări şi salvează un bebeluş. Ea ar putea trage concluzia că obiectivul pompierului a fost să salveze bebeluşul şi că principiile sale etice îl fac să pună viaţa bebeluşului mai presus decât confortul de a lenevi în maşina de pompieri, asumându-şi riscul de a se arunca în flăcări. Ar putea însă la fel de bine trage concluzia că pompierul îngheţa şi tânjea după căldură, sau că a făcut-o ca pe un exerciţiu. Dacă acest unic exemplu e tot ce ştie IA despre pompieri, incendii şi bebeluşi, i-ar fi într-adevăr imposibil să-şi dea seama care explicaţie este cea corectă. O idee de bază a învăţării prin consolidare inversă este însă că noi luăm mereu decizii, iar fiecare decizie dezvăluie ceva despre obiectivele noastre. Putem spera aşadar ca, observând o mulţime de oameni într-o mulţime de situaţii (din realitate sau din filme şi cărţi), l A să-şi construiască în cele din urmă un model exact al tuturor preferinţelor noastre.[4]

Chiar dacă poate fi construită o IA care să afle ce obiective ai, nu rezultă neapărat că le va şi adopta. Gândeşte-te la politicienii pe care îi deteşti: ştii ce vor, numai că *tu* nu vrei acelaşi lucru, şi oricât s-ar strădui, nu vor reuşi să te convingă să le adopţi obiectivele.

Avem numeroase strategii de a insufla obiectivele noastre copiilor noştri – unele mai eficiente, altele mai puţin eficiente, după cum mi-am dat şi eu seama crescând doi băieţi adolescenţi. Când trebuie să convingem calculatoare, iar nu oameni, avem de-a face cu *problema încărcării valorilor (value-loading problem),* o problemă încă mai dificilă decât educaţia morală a copiilor. Să ne gândim la un sistem LA a cărui inteligenţă e treptat ameliorată de la subuman la supraumân, mai întâi prin ajustări, apoi prin autoperfec- ţionare recursivă, ca în cazul lui Prometeu. La început e mult mai slabă decât tine, aşa încât nu te poate împiedica s-o scoţi din funcţiune şi să

înlocuieşti acele părţi din software şi din date car? codifică obiectivele ei – dar asta nu-i de folos, fiindcă e încă prea stupidă pentru a *înţelege* pe deplin obiectivele tale, a căror percepere necesită un nivel de inteligenţă uman. în cele din urmă, devine mult mai deşteaptă decât tine, şi e de sperat să poată înţelege perfect obiectivele tale – dar pesemne că nici asta nu-i de folos, fiindcă de acum e mult mai puternică decât tine şi s-ar putea să nu te lase s-o scoţi din funcţiune şi să-i înlocuieşti obiectivele, la fel cum tu nu-i laşi pe acei politicieni să-ţi înlocuiască obiectivele cu ale lor.

Cu alte cuvinte, intervalul de timp în care poţi încărca obiectivele tale în IA e probabil mic: scurta perioadă între momentul în care e încă atât de stupidă încât să te lase s-o faci şi cel în care a devenit deja atât de deşteaptă încât să nu te mai lase. Motivul pentru care încărcarea valorilor poate fi mai dificilă în cazul maşinilor decât în cazul oamenilor este că inteligenţa maşinilor sporeşte mult mai repede: la copii, intervalul acela magic de persuasiune, când inteligenţa lor e comparabilă cu cea a părinţilor, durează ani de zile, însă IA, asemenea lui Prometeu, poate depăşi acest interval în zile sau chiar ore.

Unii cercetători urmează o altă abordare pentru a înzestra maşinile cu obiectivele noastre, pe baza *corigibilităţii.* Ideea ar fi de a înzestra o IA primitivă cu un sistem de obiective aşa încât pur şi simplu să nu-i pese dacă se întâmplă s-o scoţi din funcţiune şi să-i modifici obiectivele. Dacă se dovedeşte posibil, ai putea lăsa în siguranţă IA să devină superinteligentă, s o pui în funcţiune, să instalezi obiectivele tale, iar dacă nu-ţi convin rezultatele, s-o scoţi din funcţiune şi să mai ajustezi obiectivele.

însă chiar dacă ai construi o IA care să înţeleagă obiectivele tale şi să le adopte, tot n-ai rezolvat problema alinierii obiectivelor: ce se întâmplă dacă obiectivele IA evoluează şi devin mai deştepte? Cum poţi fi sigur că ea *păstrează* obiectivele tale, indiferent câte autoperfecţionări

recursive suferă? Să analizăm un raţionament interesant care spune că păstrarea obiectivelor e automat garantată, şi să vedem dacă îi putem găsi punctele slabe.

Deşi nu putem prezice în detaliu ce se va întâmpla după o explozie a inteligenţei – motiv pentru care Vemon Vinge o numeşte „singularitate" fizicianul şi cercetătorul IA Steve Omohundro a susţinut într-un faimos articol din 2008 că putem totuşi prezice *anumite aspecte ale* comportamentului IA superinteligente, aproape indiferent ce obiective supreme ar avea.5 Acest raţionament a fost prezentat şi dezvoltat în cartea lui Nick Bostrom *Superintelligence.* Ideea de bază este că, indiferent de obiectivele ei supreme, acestea vor conduce la subobiecti ve predictibile. Am văzut în capitolul de faţă că obiectivul replicării conduce la subobiectivul de a mânca, ceea ce înseamnă că, deşi un extraterestru care ar *fi* observat evoluţia bacteriilor de pe Pământ cu miliarde de ani în urmă n-ar fi putut prezice toateobiectivele noastre umane, el ar fi putut prezice cu certitudine că *unul* dintre obiective va fi procurarea substanţelor nutritive. Privind spre viitor, ce subobiective ar trebui să ne aşteptăm să aibă o IA superinteligentă?

Din punctul meu de vedere, argumentul de bază este că, pentru a maximiza şansele de a-şi îndeplini obiectivele supreme, oricare ar fi acestea, o IA trebuie să urmărească subobiectivele din figura 7.2. Ar trebui nu doar să se străduiască să-şi perfecţioneze capacitatea de a-şi atinge obiectivele supreme, dar să se şi asigure că va păstra aceste obiective chiar şi după ce va deveni mai capabilă. Pare plauzibil: în fond, ai alege să ţi se facă un implant cerebral care să-ţi crească enorm IQ-ul, dacă ai ştii că te-ar face să-ţi doreşti să-i omori pe cei pe care îi iubeşti? Acest argument că o IA tot mai inteligentă îşi va păstra obiectivele supreme stă la baza ideii de IA prietenoasă, împărtăşită, printre alţii, de Eliezer Yudkowsky: în esenţă, ea spune că dacă vom reuşi să facem astfel încât LA

autoameliorantă să devină prietenoasă prin înţelegerea şi adoptarea obiectivelor noastre, atunci problema e rezolvată, fiindcă avem garanţia că va face tot ce-i stă în putinţă pentru a rămâne pe veci prietenoasă.

Oare e adevărat? Pentru a răspunde la întrebare, trebuie să examinăm şi celelalte subobiective din figura 7.2. IA îşi va maximiza, desigur, şansele de a îndeplini obiectivul ei suprem, oricare ar fi acesta, dacă îşi poate spori capacităţile, şi o poate face prin îmbunătăţirea hardware-ului, software ului[††††††††††] şi a modelului lu

[††††††††††] Folosesc expresia „îmbunătăţirea software-ului” în sensul cel mai larg, incluzând nu doar optimizarea algoritmilor, dar şi procesul raţional de luare a deciziilor, aşa încât să devină cât mai bun cu putinţă pentru atingerea obiectivelor. *(N. a.)*

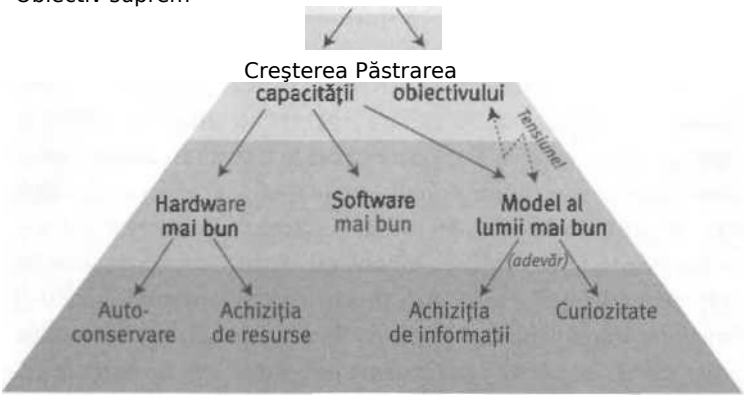


Figura 7.2 Orice obiectiv suprem al unei IA superinteligente conduce în mod firesc la subobiectiveie prezentate aici. Există însă o tensiune inerentă între păstrarea obiectivului şi îmbunătăţirea modelului pe care şi-l face despre lume, ceea ce pune sub semnul îndoielii faptul că va păstra într-adevăr obiectivul iniţial când va deveni mai deşteaptă.

mii. Acelaşi lucru e valabil la oameni: o fată al cărei obiectiv e să devină cea mai bună jucătoare de tenis din lume se va antrena să-şi îmbunătăţească hardware-ul muscular, software-ul neuronal şi modelul ei menta] despre lume care o ajută să prezică ce vor face adversarele ei. Pentru o IA, subobiectivul optimizării hardwa- re-ului ei favorizează mai buna utilizare a resurselor curente (pentru senzori, dispozitive de acţionare, calcul etc.) şi achiziţia de noi resurse. El implică de asemenea dorinţa de autoconservare, din moment ce distrugerea/defectarea ar însemna degradarea totală a hardware-ului.

Dar staţi o clipă! N u cumva cădem în capcana de a antropomorfiza IA cu această discuţie despre felul în care va încerca să adune resurse şi să se apere? Asemenea trăsături stereotipe de mascul alfa nu se întâlnesc oare numai la inteligenţele modelate de brutala evoluţie competitivă darwinistă? Odată ce inteligenţele artificiale sunt rezultatul proiectării, nu al evoluţiei, n-ar putea fi ele lipsite de ambiţie şi dotate cu spirit de sacrificiu?

Ca un studiu de caz simplu, să considerăm robotul IA din figura 7.3, al cărui unic obiectiv e să salveze de lup cât mai multe oi. Pare un obiectiv nobil şi altruist, fără legătură cu autoconservarea şi procurarea de resurse. Dar care este cea mai bună strategie pentru prietenul nostru robot? Robotul nu mai salvează nici o oaie dacă nimereşte peste bombă, deci are tot interesul să nu sară în aer. Cu alte cuvinte, îşi dezvoltă ca subobiectiv autoconservarea! Este de asemenea stimulat să manifeste curiozitate, îmbunătăţindu-şi modelul lumii prin explorarea mediului, fiindcă deşi drumul pe care merge îl va duce până la urmă pe păşune, există o scurtătură care i-ar lăsa lupului mai puţin timp să înfulece oi. în fine, dacă robotul explorează cu atenţie mediul, va descoperi importanţa procurării de resurse: poţiunea îl face să alerge mai repede, iar cu pistolul îl poate împuşca pe lup. în concluzie, nu putem considera

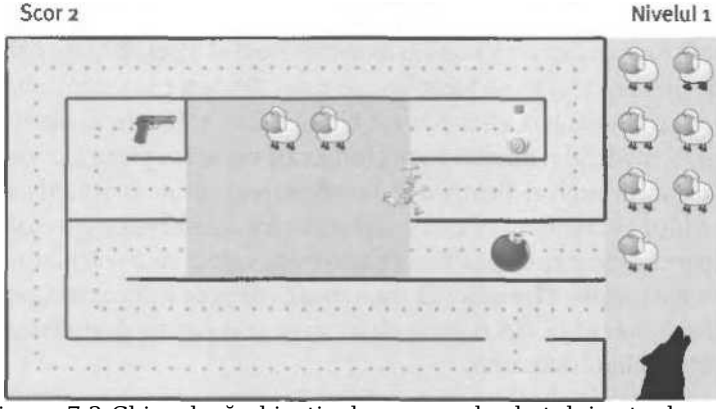


Figura 7.3 Chiar dacă obiectivul suprem al robotului este doar de a maximiza scorul aducând oile de pe păşune în staul înainte ca lupul să Ie mănânce, acesta poate conduce la subobiectivele autoconservării (evitarea bombei), explorării (găsirea unei scurtături) şi procurării de resurse (pofiunea îl face să alerge mai repede, iar cu pistolul îl poate împuşca pe lup).

că subobiective de tip „mascul alfa“, precum autoconservarea şi achiziţia de resurse, sunt relevante

doar pentru organismele care au evoluat, fiindcă robotul nostru IA le-a dezvoltat pornind de la unicul lui obiectiv de a salva oi.

Dacă dotezi o IA superinteligenţă cu unicul obiectiv de a se autodistruge, o va face fără probleme. Ea se va opune însă scoaterii din funcţiune dacă îi dai orice obiectiv care îi cere să rămână operaţională pentru a-l îndeplini – ceea ce e valabil pentru aproape toate obiectivele! Dacă îi dai unei superinteligenţe unicul obiectiv de a minimiza nenorocirile omenirii, de pildă, se va apăra împotriva încercărilor de a o scoate din funcţiune, fiindcă ştie că, în absenţa ei, ne vom face mai mult rău unii altora prin războaie şi alte nebunii.

în mod asemănător, aproape toate obiectivele pot fi mai bine îndeplinite cu mai multe resurse, aşa încât este de aşteptat ca o superinteligenţă să-şi dorească resurse, aproape indiferent de obiectivul suprem pe care îl are. A da unei superinteligenţe un obiectiv unic cu termen nelimitat şi fără constrângeri poate fi deci periculos: dacă creăm o superinteligenţă al cărei singur obiectiv este să joace cât mai bine go, soluţia raţională pentru ea este să rearanjeze sistemul nostru solar într-un calculator gigantic, fără să-i pese de foştii lui locuitori, iar apoi să înceapă colonizarea cosmosului în căutarea unei puteri de calcul mai mare. Am străbătut astfel întregul ciclu: la fel cum obiectivul de a achiziţiona resurse le-a dat oamenilor subobiectivul de a excela în jocul de go, acest obiectiv de a excela în jocul de go poate conduce la subobiectivul de a achiziţiona resurse. în concluzie, aceste subobiective emergente ne obligă să nu lăsăm liberă superinteligenţă înainte de a rezolva problema alinierii obiectivelor: dacă nu reuşim s-o înzestrăm cu obiective prietenoase faţă de oameni, e probabil ca lucrurile să ia o întorsătură rea pentru noi.

Putem aborda acum partea a treia, şi cea mai spinoasă, a problemei alinierii obiectivelor: dacă reuşim să facem ca

o superinteligenţă autoameliorantă să *înţeleagă* şi să adopte obiectivele noastre, le va *păstra* ea oare, după cum susţine Omohundro? Ce dovadă am avea?

Inteligenţa oamenilor sporeşte mult pe măsură ce cresc, dar ei nu-şi păstrează întotdeauna obiectivele din copilărie. Dimpotrivă, adesea oamenii îşi schimbă radical obiectivele atunci când descoperă lucruri noi şi devin mai înţelepţi. Câţi adulţi care urmăresc pasionaţi *Teletubbies*‡‡‡‡‡‡‡‡ cunoşti? Nu există vreo dovadă că asemenea evoluţie a obiectivelor se opreşte la un anumit prag al inteligenţei – ba chiar avem indicii că tendinţa de a înlocui obiectivele, ca reacţie la noi experienţe şi idei, nu scade, ci creşte odată cu inteligenţa.

De ce se întâmplă asta? Să ne întoarcem la subobiectivul de a construi un mai bun model al lumii – acolo e punctul nevralgic! Există o tensiune între conceperea unor modele ale lumii şi păstrarea obiectivelor (vezi figura 7.2). Sporirea inteligenţei poate aduce cu sine nu numai o îmbunătăţire cantitativă a capacităţii de a atinge aceleaşi vechi obiective, dar şi o înţelegere calitativ diferită a naturii realităţii, care arată că vechile obiective sunt nechibzuite, absurde sau chiar nedefinite. De pildă, să presupunem că programăm o IA prietenoasă pentru a maximiza numărul de oameni ale căror suflete ajung în rai după moarte. Mai întâi, aceasta încearcă, de pildă, să sporească compasiunea oamenilor şi participarea la slujbele religioase. Să presupunem însă că apoi ajunge la o înţelegere ştiinţifică a oamenilor şi a conştiinţei umane, şi, spre marea ei surprindere, descoperă că sufletul nu există. Ce-i de făcut? În mod asemănător, e posibil ca IA să descopere în cele din urmă că orice alt obiectiv pe care i-l dăm pe baza înţelegerii noastre actuale a lumii (cum ar fi „maximizarea sensului vieţii umane”) este nedefinit.

Apoi, în tentativa de a concepe modele mai bune ale

‡‡‡‡‡‡‡‡‡‡ Serial TV britanic pentru preşcolari. (JV.t)

lumii, IA ar putea căuta, la fel cum am făcut-o noi, modele pentru sine, şi ar putea încerca să înţeleagă cum funcţionează – cu alte cuvinte, ar reflecta asupra ei înseşi. Odată ce va construi un bun model al ei şi îşi va da seama ce este ea, va înţelege obiectivele pe care i le-am dat la un meta-nivel, şi va alege pesemne să le ignore sau să le saboteze, la fel cum noi, oamenii, înţelegem şi sabotăm în mod deliberat obiectivele pe care genele ni le-au dat – apelând, de exemplu, la controlul natalităţii. Am văzut deja în secţiunea despre psihologie de ce alegem să ne păcălim genele şi să le sabotăm obiectivul: fiindcă suntem loiali doar amestecului nostru de preferinţe afective, iar nu obiectivului genetic care le-a creat – ceea ce acum înţelegem şi ni se pare banal. Alegem aşadar să ne deturnăm mecanismul de recompense, exploatându-i punctele slabe. în mod asemănător, obiectivul de protejare a valorilor umane pe care îl programăm în IA prietenoasă devine codul genetic al maşinii. Odată ce această IA prietenoasă se înţelege pe sine suficient de bine, poate considera că obiectivul e banal sau nechibzuit, aşa cum ni se pare nouă reproducerea compulsivă, şi nu-i deloc evident că nu va găsi o cale de a-l sabota, exploatând punctele slabe din programarea noastră.

De exemplu, să presupunem că un grup de furnici te creează ca să fii un robot autoameliorant, mult mai deştept decât ele, care le împărtăşeşte obiectivele şi le ajută să construiască muşuroaie tot mai mari, iar în cele din urmă ajungi la nivelul de inteligenţă uman la care te afli acum. Crezi că-ţi vei petrece restul zilelor doar optimizând muşuroaie, sau vei fi atras de probleme şi scopuri mai sofisticate, imposibil de înţeles pentru furnici? în al doilea caz, crezi că vei găsi o cale să ignori pornirea de a proteja furnicile cu care te-au înzestrat creatorii tăi, la fel cum tu, cel din realitate, ignori unele porniri cu care te-au înzestrat genele? Atunci, o IA prietenoasă superinteligentă

nu ar putea oare considera că obiectivele noastre umane actuale sunt la fel de banale şi insipide cum ţi se par cele ale furnicilor, şi n-ar elabora oare noi obiective, diferite de cele pe care Ie-a învăţat şi le-a adoptat de la noi?

Nu-i exclus să existe un mod de a proiecta o IA autoamelio- rantă care să ne dea garanţia că păstrează pe veci obiective prietenoase faţă de oameni, dar trebuie să recunosc că deocamdată nu ştim cum s-o construim – şi nici măcar dacă e posibil. în concluzie, problema alinierii obiectivelor IA are trei părţi; nici una nu e rezolvată, iar toate sunt în prezent subiect de cercetare intensă. Fiind atât de dificile, ar trebui să Ie consacram eforturile noastre încă de-acum, cu mult înainte de apariţia vreunei super-inteligenţe, pentru a fi siguri că vom avea răspunsurile la momentul oportun.

ETICĂ: ALEGEREA OBIECTIVELOR

Am discutat până acum despre înţelegerea, adoptarea şi păstrarea obiectivelor noastre de către maşini. Dar cine suntem „noi“? Despre ale cui obiective vorbim? Ar trebui oare o persoană sau un grup să hotărască ce obiective va adopta o superinteligenţă viitoare, în ciuda faptului că sunt diferenţe uriaşe între obiectivele lui Adolf Hitler, Papa Francisc şi Cari Sagan? Sau există unele obiective consensuale, reprezentând un compromis acceptabil pentru omenire în ansamblul ei?

Cred că ambele probleme, şi cea etică, şi cea a alinierii obiectivelor, sunt probleme cruciale care trebuie rezolvate înainte de crearea unei superinteligenţe. Pe de o parte, amânarea lămuririi aspectelor etice până după ce se construieşte o superinteligenţă cu obiective aliniate ar fi iresponsabilă, ba chiar catastrofală. O superinteligenţă perfect obedientă ale cărei obiective se aliniază automat cu

cele ale proprietarului ei uman ar fi un fel de Obersturmbannführer SS Adolf Eichmann mult amplificat: lipsită de busolă morală şi de inhibiţii, ar pune în aplicare cu o eficienţă nemiloasă obiectivele proprietarului ei, oricare ar fi acestea.[6] Pe de altă parte, doar dacă rezolvăm problema alinierii obiectivelor are rost să aducem în discuţie obiectivele care ar trebui alese – ceea ce vă invit să facem acum.

încă din Antichitate filozofii au visat să deducă etica (principiile care ne spun cum ar trebui să ne comportăm) pornind de la zero, folosind doar principii incontestabile şi logica. Mii de ani mai târziu însă, singurul consens la care au ajuns este că nu există vreun consens. De pildă, în vreme ce Aristotel punea accentul pe virtuţi, Immanuel Kant punea accentul pe îndatoriri, iar utilitariştii pe fericirea cea mai mare pentru cei mai mulţi. Kant susţinea că putea deduce din principii prime (numite de el „imperative categorice") concluzii pe care numeroşi filozofi contemporani nu le împărtăşesc: masturbarea e mai rea decât sinuciderea, homosexualitatea e condamnabilă, bastarzii pot fi ucişi, iar soţiile, servitorii şi copiii au statut de obiect pentru proprietarii lor.

Pe de altă parte, în ciuda acestui dezacord, există multe teme etice în privinţa cărora, indiferent de cultură sau de epocă, oamenii au căzut de acord. De pildă, accentul pus *pe frumos, bine* şi *adevăr* apare şi în *Bhagavad Gita,* şi la Platon. Institutul pentru Studii Avansate de la Princeton, unde am lucrat ca postdoc, are deviza „Adevăr & Frumos", în vreme ce deviza Universităţii Harvard, lăsând deoparte esteticul, se rezumă la „Veritas", adevăr. în cartea sa *A Beautiful Question,* colegul meu Frank Wilczek susţine că adevărul e legat de frumos şi că putem privi universul nostru ca pe o operă de artă. Ştiinţa, religia şi filozofia năzuiesc toate trei spre adevăr. Religia pune un puternic accent asupra binelui, la fel şi universitatea mea, MIT: în discursul inaugural din 2015, rectorul Rafael Reif spunea

că misiunea noastră este să sporim binele în lume.

Deşi încercările de a deduce o etică unanim acceptată pornind de la zero au eşuat, e larg împărtăşită ideea că unele principii etice decurg din principii fundamentale, ca subobiective ale unor obiective fundamentale. De exemplu, năzuinţa spre adevăr poate fi privită drept căutarea unui model mai bun al lumii (figura 7.2): pe înţelegerea naturii ultime a realităţii se bazează unele obiective etice. Dispunem acum de un cadru excelent pentru căutarea adevărului – metoda ştiinţifică Dar cum putem determina ce e frumos sau ce e bun? Anumite aspecte ale frumosului pot fi şi ele regăsite în obiective de nivel mai profund. De pildă, standardele noastre de frumuseţe masculină şi feminină ar putea reflecta parţial evaluări subconştiente ale şanselor de replicare pentru genele noastre.

Cât priveşte binele, aşa-numita Regulă de Aur (poartă-te cu alţii aşa cum ai vrea să se poarte ei cu tine) apare in majoritatea culturilor şi religiilor, şi urmăreşte în mod evident să încurajeze perpetuarea armonioasă a societăţii umane (deci şi a genelor noastre) prin întărirea colaborării şi descurajarea dezbinării sterile.[7] Acelaşi lucru e valabil şi pentru alte reguli etice mai explicite care au fost aşezate la loc de cinste în sisteme juridice din lumea întreagă, cum ar fi accentul pus de confucianism pe cinste şi multe dintre cele Zece Porunci, între care „Să nu ucizi“. Cu alte cuvinte, numeroase principii etice se leagă de emoţii sociale cum sunt empatia şi compasiunea: ele au evoluat pentru a genera colaborarea şi ne influenţează comportamentul prin recompense şi pedepse. Dacă facem un lucru rău şi îi regretăm consecinţele, pedeapsa noastră afectivă e aplicată direct de chimia creierului. Pe de altă parte, dacă încălcăm principiile etice, societatea ne poate pedepsi pe căi indirecte, cum fi oprobriul public sau condamnarea pentru nerespectarea legilor.

Aşadar, deşi astăzi omenirea nu a ajuns la un consens în

privinţa eticii, există numeroase principii de bază larg împărtăşite. Şi nu e de mirare, din moment ce societăţile umane care au supravieţuit până în prezent au în general principii etice care au fost optimizate pentru acelaşi obiectiv: încurajarea supravieţuirii şi prosperităţii lor. Privind spre viaţa din viitor care are potenţialul de a prospera în cosmos vreme de miliarde de ani, asupra cărui set minim de principii etice am putea cădea de acord că trebuie respectat? Iată o dezbatere la care e necesar să participăm cu toţii. M-a fascinat să ascult şi să citesc opiniile a numeroşi gânditori, şi cred că majoritatea ideilor lor pot fi distilate în patru principii:

- Utilitarism: Experienţele pozitive ale conştiinţei trebuie maximizate, iar suferinţa trebuie minimizată.
- Diversitate: Un set divers de experienţe pozitive e preferabil repetiţiilor numeroase ale aceleiaşi experienţe, chiar dacă aceasta din urmă a fost identificată drept cea mai bună dintre experienţele pozitive.
- Autonomie: Entităţile/societăţile conştiente trebuie să aibă libertatea de a-şi urmări propriile obiective, cu condiţia să nu intre în contradicţie cu un principiu de ordin superior.
- Moştenire: Compatibilitate cu scenariile pe care *astăzi* majoritatea oamenilor le-ar considera fericite, incompatibilitate cu scenariile pe care aproape toţi oamenii le-ar considera *astăzi* cumplite.

Să zăbovim puţin asupra acestor patru principii. Tradiţional, se consideră că utilitarism înseamnă „fericirea cea mai mare pentru cel mai mare număr de oameni“, dar eu l-am generalizat aici pentru a fi mai puţin antropocentric, aşa încât să poată include şi animale non-umane, minţi umane simulate conştiente şi alte IA care pot apărea în viitor. Am folosit în definiţie termenul *experienţe,* nu oameni sau lucruri, fiindcă majoritatea gânditorilor sunt de acord că senzaţia de frumos, bucuria,

plăcerea şi suferinţa sunt experienţe subiective. Prin urmare, dacă nu există experienţe (ca într-un univers mort sau intr-unui populat cu maşini inconştiente de tip zombi), nu poate exista vreun sens şi nimic din ceea ce este etic relevant. Dacă acceptăm acest principiu utilitarist, este esenţial să ne dăm seama ce sisteme inteligente sunt conştiente (în sensul că au o experienţă subiectivă) şi ce sisteme nu sunt conştiente – despre asta vom vorbi în următorul capitol.

Dacă acest principiu utilitarist ar fi singurul de care am ţine cont, atunci pesemne că am vrea să aflăm care este cea mai bună dintre experienţele pozitive posibile, iar apoi am coloniza cosmosul şi am re crea exact aceeaşi experienţă (şi nimic altceva) iarăşi şi iarăşi, de câte ori e cu putinţă, în cât mai multe galaxii – folosind simulări, dacă aceasta e modalitatea cea mai eficientă. Dacă ţi se pare un mod banal de a ne folosi zestrea cosmică, bănuiesc că măcar o parte din ceea ce simţi că lipseşte acestui scenariu este diversitatea. Cum ţi s-ar părea să mănânci toată viaţa mereu ace laşi lucru? Să vezi mereu acelaşi film? Toţi prietenii tăi să arate identic, să aibă personalităţi şi idei identice? Poate că o parte din preferinţa noastră pentru diversitate ţine de faptul că diversitatea a ajutat omenirea să supravieţuiască şi să prospere, făcându-ne mai robuşti. Poate că e legată şi de preferinţa pentru inteligenţă: sporirea inteligenţei în cursul celor 13,8 miliarde de ani de istorie cosmică a transformat uniformitatea anostă în structuri tot mai diverse, diferenţiate şi complexe care procesează informaţia în moduri tot mai complicate.

Principiul autonomiei se află la baza a numeroase drepturi şi libertăţi stipulate în Declaraţia Universală a Drepturilor Omului, adoptată de Naţiunile Unite în 1948 în încercarea de a trage învăţăminte de pe urma celor două războaie mondiale. Aceasta se referă la libertatea de

gândire, exprimare şi mişcare, interzicerea sclaviei şi torturii, dreptul la viaţă, libertate, securitate şi educaţie, dreptul la fericire, muncă şi proprietate. Dacă vrem să fim mai puţin antropocentrici, putem vorbi în termeni mai generali despre libertatea de a gândi, învăţa, comunica, deţine proprietăţi, precum şi de dreptul de a face orice nu încalcă libertăţile celorlalţi. Principiul autonomiei susţine diversitatea, căci nu toţi împărtăşesc exact aceleaşi obiective. Apoi, acest principiu al autonomiei decurge din principiul utilitarist, dacă entităţile individuale au drept obiective experienţele pozitive şi încearcă să acţioneze în propriul lor interes: presupunând că, dimpotrivă, am interzice unei entităţi să-şi urmărească obiectivul, deşi n-ar provoca vreun neajuns nimănui, ar exista în total mai puţine experienţe pozitive. Este tocmai argumentul folosit de economişti în favoarea pieţei libere: aceasta conduce în mod natural la o situaţie eficientă (numită de economişti „optimalitate Pareto“) în care nimeni n-o poate duce mai bine fără ca un altul s-o ducă mai rău.

Principiul moştenirii afirmă în esenţă că ar trebui să avem ceva de spus despre viitor, din moment ce contribuim la edificarea lui. Principiile autonomiei şi moştenirii întruchipează idealuri democratice: primul dă vieţii viitoare puterea de a decide cum va fi folosită zestrea cosmică, în vreme ce al doilea le dă chiar oamenilor de azi ceva din această putere de decizie.

Deşi cele patru principii nu par să genereze controverse, ele sunt greu de pus în practică, fiindcă diavolul se ascunde în detalii. Dificultatea ne aminteşte de celebrele „Trei legi ale roboticii** imaginate de Isaac Asimov:

r. Un robot n-are voie să pricinuiască vreun rău unei fiinţe umane, sau, prin neintervenţie, să permită ca unei fiinţe umane să i se facă un rău.
2. Un robot trebuie să se supună ordinelor date de o

fiinţă umană, atâta vreme cât ele nu intră în contradicţie cu prima lege.

3. Un robot trebuie să-şi protejeze propria existenţă, atâta vreme cât această protecţie nu intră în contradicţie cu prima sau cu a doua lege.

Deşi totul pare să fie în regulă, multe dintre povestirile lui Asimov arată că aceste legi conduc la contradicţii problematice în situaţii neprevăzute. Să presupunem acum că înlocuim aceste trei legi cu două, în încercarea de a codifica principiul autonomiei pentru formele de viaţă din viitor;

i. O entitate conştientă e liberă să gândească, să înveţe, să comunice, să deţină proprietăţi, şi nu e voie să fie vătămată sau distrusă.
2. 0 entitate conştientă are dreptul să facă orice nu intră în contradicţie cu prima lege.

Sună bine, nu-i aşa? Dar staţi să ne gândim o clipă. Dacă animalele sunt conştiente, atunci ce ar trebui să mănânce prădătorii? Toţi prietenii noştri sunt oare obligaţi să devină vegetarieni? Dacă nişte programe de calculator sofisticate se vor dovedi a fi conştiente, va fi ilegal să le oprim? Dacă există reguli împotriva lichidării formelor de viaţă digitale, ar trebui oare să existe şi restricţii privind crearea lor, pentru a evita o explozie a populaţiei digitale? Declaraţia Universală a Drepturilor Omului s-a bucurat de un larg consens pentru că doar oamenii au fost întrebaţi. De îndată ce luăm în considerare un spectru mai larg de entităţi conştiente, cu grade diferite de înzestrare şi putere, suntem confruntaţi cu compromisuri dificile între protejarea celui slab şi ideea că „dreptatea este de partea celui mai tare“.

Probleme delicate apar şi în legătură cu principiul moştenirii. Dată fiind evoluţia perspectivelor etice asupra

sclaviei, drepturilor femeilor etc. din Evul Mediu încoace, am vrea oare ca oamenii de acum r 500 de ani să influenţeze masiv felul în care e guvernată lumea de azi? Dacă nu, de ce să încercăm să impunem etica noastră fiinţelor din viitor, care ar putea fi mult mai deştepte decât noi? Chiar credem că o IAG supraumană şi-ar dori ceea ce preţuieşte intelectul nostru inferior? Ar fi ca şi cum un copil de patru ani şi-ar închipui că atunci când va creşte mare şi va deveni mult mai deştept îşi va dori să construiască o uriaşă casă din turtă dulce în care să-şi petreacă toată ziua mâncând bomboane şi îngheţată. Aidoma copilului, viaţa de pe Pământ e probabil să depăşească preocupările ei infantile. Sau închipuiţi-vă un şoarece care creează o IAG la nivel uman şi crede că aceasta va vrea să construiască oraşe întregi din brânză. Pe de altă parte, dacă am şti că o IA supraumană ar putea ajunge într-o bună zi la cosmocid, distrugând întreaga viaţă din universul nostru, de ce ar accepta oamenii de azi un viitor lipsit de viaţă, dacă 1 ar putea împiedica făurind altfel inteligenţa artificială de mâine?

în concluzie, este extrem de dificil să codifici în întregime chiar şi principii etice larg acceptate în forme aplicabile viitoarei IA, iar această problemă necesită discuţii şi studii serioase, pe măsură ce LA progresează. Dar între timp să nu lăsăm ca perfecţiunea să devină inamicul binelui: există numeroase exemple de „etică de grădiniţă" necontroversată care pot şi trebuie să fie încorporate în tehnologia de mâine. De pildă, avioanele de pasageri n-ar trebui să se poată izbi de obiecte staţionare, iar acum când practic toate au pilot automat, radar şi GPS, nu mai există scuze de ordin tehnic. Şi totuşi, pe n septembrie 2001, piraţi ai aerului au izbit trei aeronave de clădiri, iar pe 24 martie 2015 pilotul sinucigaş Andreas Lubitz a îndreptat avionul companiei German wings spre un munte – fixând pilotul automat la altitudinea de 30 de

metri deasupra nivelului mării şi lăsând computerul de zbor să se ocupe de restul. Acum când maşinile noastre sunt suficient de deştepte pentru a avea unele informaţii despre ceea ce fac, e momentul să le învăţăm să-şi impună limite. Orice inginer care proiectează o maşină trebuie să se întrebe dacă sunt lucruri pe care maşina le poate face, dar n-ar trebui să le facă, şi să se gândească dacă există o modalitate practică de a împiedica un utilizator răuvoitor sau neîndemânatic să provoace nenorociri.

OBIECTIVE SUPREME?

Acest capitol a fost o scurtă istorie a obiectivelor noastre. Dacă am putea privi derulat la viteză mare filmul celor 13,8 miliarde de ani de istorie cosmică, am distinge câteva stadii de comportament orientat spre obiective:

1. Materia părând să urmărească maximizarea *disipării ei.*
2. Viaţa primitivă părând să încerce să-şi maximizeze *replicarea.*
3. Oameni urmărind nu replicarea, ci obiective legate de plăcere, curiozitate, compasiune şi alte senzaţii şi sentimente pe care le-au dezvoltat pentru a-i ajuta la replicare.
4. Maşini construite de oameni pentru a-i ajuta să-şi urmărească obiectivele.

Dacă maşinile vor declanşa în cele din urmă o explozie a inteligenţei, cum se va sfârşi această istorie a obiectivelor? Ar putea exista un sistem de obiective sau un cadru etic către care să conveargă aproape toate entităţile, pe măsură ce devin din ce în ce mai inteligente? Cu alte cuvinte, avem oare un soi de destin etic?

O citire superficială a istoriei omenirii ar putea sugera indicii privind o asemenea convergenţă: în cartea sa *The BetterAngels of Our Nature,* Steven Pinker susţine că, în decursul a mii de ani, omenirea a devenit mai puţin

violentă şi mai cooperantă, iar în multe părţi ale lumii a crescut acceptarea diversităţii, autonomiei şi democraţiei. Apoi, în ultimele milenii, căutarea adevărului prin metode ştiinţifice s-a răspândit din ce în ce mai mult. S-ar putea însă ca aceste tendinţe să indice convergenţa subobiectivelor, nu a obiectivelor supreme. De exemplu, în figura 7.2, căutarea adevărului (un model mai precis al lumii) apare doar ca subobiec- tiv al aproape oricărui obiectiv suprem. în mod asemănător, am văzut mai sus că principii etice precum cooperarea, diversitatea şi autonomia pot fi privite ca subobiective, în sensul că ele ajută societăţile să funcţioneze eficient, deci să supravieţuiască şi să îndeplinească acele obiective fundamentale pe care acestea le au. Unii ar putea susţine chiar că tot ce numim „valori umane“ este doar un protocol de cooperare, făcând ca subobiectivul cooperării să funcţioneze mai eficient. Privind spre viitor şi urmând aceeaşi idee, este probabil ca orice IA superinteligentă să aibă între sub- obiective un hardware mai eficient, un software mai eficient, căutarea adevărului şi curiozitatea, pur şi simplu fiindcă aceste subobiective o ajută să-şi îndeplinească obiectivele supreme, indiferent care ar fi ele.

în cartea sa *Superintelligence,* Nick Bostrom contrazice ipoteza destinului etic, prezentând ceea ce el numeşte *teza ortogonalităţii:* obiectivele supreme ale unui sistem pot fi independente de inteligenţa lui. Prin definiţie, inteligenţa este doar capacitatea de a îndeplini obiective complexe, indiferent de natura lor, aşa încât teza ortogonalităţii pare rezonabilă. în fond, oamenii pot fi inteli- genţi şi buni sau inteligenţi şi cruzi, iar inteligenţa poate fi folosită pentru a face descoperiri ştiinţifice, a crea artă, a-i ajuta pe oameni, sau pentru a plănui atacuri teroriste.[8]

Teza ortogonalităţii are meritul de a ne spune că obiectivele supreme ale vieţii în cosmosul nostru nu sunt predestinate, ci avem libertatea şi capacitatea de a le modela. Ea sugerează că acea convergenţă către un

obiectiv unic nu ţine de viitor, ci de trecut, când întreaga viaţă a apărut având ca unic obiectiv replicarea. Odată cu trecerea timpului cosmic, minţi tot mai inteligente ajung să se revolte, să se elibereze de obiectivul banal al replicării şi să-şi aleagă propriile lor obiective. Noi, oamenii, nu suntem pe deplin liberi în acest sens, căci multe obiective rămân genetic înscrise în noi, dar inteligenţele artificiale se pot descătuşa complet de obiectivele anterioare. Această libertate sporită în ce priveşte obiectivele este vizibilă azi la sistemele IA limitate: după cum am arătat mai sus, singurul obiectiv al unui computer pentru jocul de şah este să câştige la şah, dar există şi computere al căror obiectiv este să piardă la şah, şi care participă la turnee în care obiectivul e să-ţi obligi adversarul să-ţi captureze piesele. Pesemne că această eliberare de preferinţele induse de evoluţie poate face ca inteligenţele artificiale să fie, într-un sens profund, mai etice decât oamenii: filozofi ai moralei precum Peter Singer susţin că în genere oamenii se comportă neetic din cauze evolutive – discriminând, de pildă, animalele non-umane.

Am văzut că la baza scenariului „IA prietenoasă“ stă ideea că o IA recursiv autoameliorantă îşi va dori să păstreze obiectivul ei suprem (prietenos) pe măsură ce devine tot mai inteligentă. Dar cum ar putea fi măcar definit un „obiectiv suprem“ (sau „obiectiv final“, după expresia lui Bostrom) pentru o superinteligenţă? Cred că putem avea încredere în scenariul „IA prietenoasă” doar cu condiţia să răspundem la această întrebare crucială.

În cercetarea IA, maşinile inteligente au de regulă un obiectiv final clar, bine definit – de pildă, de a câştiga la şah sau a conduce la destinaţie un automobil, în condiţii legale. Acelaşi lucru e valabil pentru majoritatea sarcinilor pe care le încredinţăm oamenilor, fiindcă orizontul de timp şi contextul sunt cunoscute şi limitate. Acum vorbim însă despre întregul viitor al vieţii în universul nostru, limitat

doar de legile fizicii (încă necunoscute pe deplin), aşa încât definirea unui obiectiv e o misiune copleşitoare! Lăsând deoparte efectele cuantice, un obiectiv cu adevărat bine definit ar trebui să indice cum trebuie aranjate toate particulele din universul nostru la sfârşitul timpului. Dar nu-i deloc limpede dacă există un sfârşit al timpului bine definit în fizică. Dacă particulele sunt aranjate în acel mod la un moment anterior, aranjamentul are puţine şanse de a dăinui. Iar apoi, ce aranjament al particulelor e preferabil?

Oamenii tind să prefere anumite aranjamente ale particulelor în detrimentul altora; de exemplu, preferăm ca oraşul nostru să fie aranjat aşa cum e acum, în loc să fie rearanjat de explozia unei bombe cu hidrogen. Să presupunem că am vrea să definim *o funcţie a calităţii* care asociază un număr fiecărui aranjament posibil al particulelor din universul nostru, cuantificând cât de „bun" credem că este aranjamentul, iar apoi să trasăm unei IA superin- teligente obiectivul de a maximiza această funcţie. Ar putea părea o abordare rezonabilă, din moment ce descrierea comportamentului orientat spre obiective ca maximizarea unei funcţii e întâlnită şi în alte domenii ale ştiinţei: de exemplu, economiştii îşi construiesc modele conform cărora oamenii încearcă să maximizeze ceea ce ei numesc „funcţie de utilitate", iar mulţi proiectanţi LA îşi antrenează agenţii inteligenţi pentru a maximiza ceea ce ei numesc „funcţie de recompensă". Când vorbim însă despre obiective supreme pentru cosmosul nostru, această abordare devine un coşmar computaţional, fiindcă ar trebui definită o valoare a calităţii pentru fiecare din acel peste un googolplex de aranjamente posibile ale particulelor elementare din universul nostru – un googolplex este i urmat de io^{100} zerouri, mai multe zerouri decât particulele din universul nostru. Cum am putea defini această funcţie a calităţii pentru o IA?

După cum am văzut mai sus, singurul motiv pentru care

noi, oamenii, avem preferinţe ar putea fi acela că suntem soluţia unei probleme evolutive de optimizare. Prin urmare, toate cuvintele cu valoare normativă din limbajul nostru uman, cum ar fi „delicios“, „parfumat“, „frumos“, „confortabil“, „interesant“, „sexy“, „semnificativ“, „fericit“ şi „bun“, îşi au originea în această optimizare evolutivă: nu avem aşadar garanţia că o IA superinteligentă le-ar considera riguros definisabile. Chiar dacă IA ar învăţa să prezică cu exactitate preferinţele unor oameni reprezentativi, n-ar putea calcula funcţia calităţii pentru cea mai mare parte a aranjamentelor de particule: imensa majoritate a aranjamentelor posibile corespund unor scenarii cosmice stranii, fără stele, planete sau oameni, legat de care oamenii n-au nici o experienţă, aşa încât cine poate spune cât de „bune” sunt ele?

Desigur, există *unele* funcţii ale aranjamentului particulelor cosmice care pot fi riguros definite, şi cunoaştem sisteme fizice care evoluează pentru a le maximiza. De pildă, am văzut deja că multe sisteme evoluează pentru a-şi maximiza *entropia,* care în absenţa gravitaţiei conduce la moartea termică, în care totul e dezolant de uniform şi neschimbător. Entropia nu este deci un lucru pe care am vrea ca IA să-l considere „bun“ şi să se străduiască să-l maximizeze. Iată câteva exemple de cantităţi pe care am putea încerca să le maximizăm şi care ar putea fi riguros definite în termeni de aranjamente ale particulelor:

- Fracţiunea din întreaga materie a universului nostru care ia forma unui anumit organism – să zicem oameni sau *E. coli* (inspiraţi de maximizarea evolutivă a valorii selective inclusive)[§§§§§§§§§§]
- Capacitatea unei IA de a prezice viitorul, considerată

§§§§§§§§§§ Valoarea selectivă inclusivă *(inclusivejîtness)* reprezintă numărul de progenituri pe care un individ le creşte, indiferent cine le-a procreat, *(N.* t)

de Marcus Hutter o bună măsură a inteligenţei acesteia

- Ceea ce Alex Wissner-Gross şi Cameron Freer numesc *entropie cauzală* (un mandat pentru ocaziile din viitor) şi consideră a fi semnul distinctiv al inteligenţei
- Capacitatea de calcul a universului nostru
- Complexitatea algoritmică a universului nostru (de câţi biţi e nevoie pentru a-l descrie)
- Cantitatea de conştiinţă din universul nostru (vezi capitolul următor)

Privind însă din perspectiva fizicii, conform căreia cosmosul e alcătuit din particule elementare aflate în mişcare, e greu de văzut cum ar putea părea specială în mod natural o interpretare sau alta privind ceea ce e „bun". Trebuie identificat orice obiectiv final pentru universul nostru care e deopotrivă definisabil şi dezirabil. Singurele obiective programabile în prezent care vor rămâne cu certitudine bine definite pe măsură ce o IA va deveni tot mai inteligentă simt obiectivele exprimate doar în termenii mărimilor fizice, cum sunt aranjamentele particulelor, energia şi entropia. Nu avem însă nici un motiv să credem că asemenea obiective definisabile vor fi dezirabile pentru a garanta supravieţuirea omenirii

Dimpotrivă, noi, oamenii, se pare că suntem un accident istoric, iar nu soluţia optimă a unei probleme de fizică bine definite. Aceasta sugerează că o IA superinteligentă cu un obiectiv riguros definit îşi va atinge mai lesne obiectivul eliminându-ne. Ceea ce înseamnă că, pentru a lua hotărâri înţelepte în privinţa dezvoltării IA, trebuie să ne confruntăm nu numai cu provocările tradiţionale de ordin computaţional, dar şi cu întrebări filozofice dintre cele mai dificile. Pentru a programa un automobil care se conduce singur, trebuie să rezolvăm problema tramvaiului********** –

********** Forma modernă a „problemei tramvaiului" a fost enunţată în 1967 de filozoful britanic Philippa Foot. Datele

pe cine sacrificăm? Pentru a programa o IA prietenoasă, trebuie sa lămurim sensul vieţii. Ce înseamnă „sens“? Ce înseamnă „viaţă“? Care este imperativul etic ultim? Cu alte cuvinte, ce ar trebui să facem pentru a modela viitorul universului? Dacă cedăm unei superin- teligenţe controlul înainte de a răspunde riguros la aceste întrebări, răspunsul ei este puţin probabil să ne implice şi pe noi. A sosit deci momentul să ne întoarcem la dezbaterile clasice ale filozofiei şi eticii.

REZUMAT:

• Originea ultimă a comportamentului orientat spre obiective se află în legile fizicii, care implică optimizarea.

• Termodinamica încorporează obiectivul *disipării:* creşterea unei mărimi fizice care măsoară gradul de dezordine numită *entropie.*

• Viaţa e un fenomen care poate accelera disiparea (creşterea globală a dezordinii) prin menţinerea sau sporirea complexităţii ei şi prin replicare, făcând în acelaşi timp să crească dezordinea din jur.

■ Evoluţia darwinistă deturnează comportamentul orientat spre obiective de la disipare la replicare.

• Inteligenţa este capacitatea de a îndeplini obiective complexe.

• Din moment ce noi, oamenii, n-aveam întotdeauna resursele pentru a găsi strategia de replicare cu adevărat optimă, am elaborat reguli empirice utile care călăuzesc hotărârile noastre: senzaţii şi sentimente precum foamea, setea, durerea, dorinţa sexuală şi compasiunea.

• Ca atare, nu mai avem un obiectiv simplu cum e replicarea; când sentimentele şi senzaţiile noastre intră în contradicţie cu obiectivul genelor noastre, ne urmăm sentimentele şi senzaţiile, ca în cazul controlului natalităţii.

■ Construim maşini tot mai inteligente pentru a ne ajuta la îndeplinirea obiectivelor. în măsura în care construim asemenea maşini care să manifeste un comportament orientat spre obiective, ne străduim să aliniem obiectivele maşinii cu obiectivele noastre.

• Alinierea obiectivelor maşinii cu obiectivele noastre implică trei probleme nerezolvate: înţelegerea, adoptarea şi păstrarea de către

problemei sunt următoarele: un tramvai se îndreaptă spre cinci oameni legaţi de şine, iar tu te afli lângă maneta macazului şi poţi devia tramvaiul pe o linie secundară, unde unde se află legat de şine un singur om. Variantele sunt: 1) Nu faci nimic, şi tramvaiul îşi continuă drumul pe linia principală, ucigându-i pe cei cinci. *2)* Acţionezi maneta, şi tramvaiul e deviat pe linia secundară, unde ucide un om. Ce alegi să faci? *(N.t)*

maşini a acelor obiective.
• IA poate fi creată pentru a avea practic orice obiectiv, dar aproape orice obiectiv suficient de ambiţios poate conduce la subobiectivele autoconservării, achiziţiei de resurse şi curiozităţii pentru a înţelege mai bine lumea – primele două ar putea face ca o IAsuperinteligentă să provoace oamenilor dificultăţi, iar a treia ar putea-o împiedica să păstreze obiectivele pe care i le dăm.
• Deşi majoritatea oamenilor cad de acord asupra unor principii etice generale, nu e clar cum ar putea fi acestea aplicate altor entităţi precum animalele non-umane şi viitoarele inteligenţe artificiale.
• Nu e limpede cum am putea insufla unei IA superinteligente un obiectiv suprem care să fie bine definit şi să nu conducă la eliminarea omenirii, aşa încât devine imperios necesar să ne întoarcem la unele dintre cele mai spinoase probleme filozofice.

CAPITOLUL 8

Conştiinţa

Nu-mi pot închipui o teorie despre tot ce există care să ignore conştiinţa. **— Andrei Linde, 2002**

Trebuie să ne străduim să sporim conştiinţa însăţi – să proiectăm o lumină mai puternică şi mai strălucitoare într-un univers altminteri întunecat. **— Giulio Tenorii, 2012**

Am văzut că IA ne poate ajuta să făurim un viitor minunat dacă reuşim să găsim răspunsuri la unele dintre cele mai vechi şi mai grele probleme filozofice – înainte să avem nevoie de aceste răspunsuri. După cum spunea Nick Bostrom, filozofia se confruntă cu un termen-limită. Vă propun să examinăm în capitolul de faţă unul dintre cele mai delicate subiecte filozofice: conştiinţa.

CUI îl PASĂ?

Conştiinţa suscită numeroase dezbateri. Dacă pomeneşti

„cuvântul acela care începe cu C“ unui cercetător IA, unui specialist în neuroştiinţe sau unui psiholog, acesta şi-ar da pesemne ochii peste cap. Dacă se întâmplă să fie profesorul tău, te-ar putea compătimi şi ţi-ar zice că-ţi pierzi vremea cu o problemă neştiinţifică insolubilă. Prietenul meu Christof Koch, care conduce în prezent Institutul Allen pentru Ştiinţe ale Creierului, mi-a spus că, înainte de a se angaja, a fost avertizat să nu lucreze în domeniul conştiinţei de însuşi laureatul Nobel Francis Crick. în *Dicţionarul de psihologie Macmillan,* la articolul „conştiinţă“ se menţionează că „nimic demn de a fi citit n-a fost scris despre asta“.[1] După cum voi arăta în acest capitol, eu sunt mai optimist.

Deşi gânditorii au meditat la misterul conştiinţei vreme de mii de ani, apariţia IA ne impune să rezolvăm urgent probleme cum ar fi aceea de a prezice care dintre entităţile inteligente au experienţe subiective. Aşa cum am văzut în capitolul 3, putem răspunde la întrebarea dacă maşinilor inteligente trebuie să li se ofere o formă oarecare de drepturi doar în măsura în care ştim dacă ele posedă conştiinţă şi pot suferi sau simţi bucurie. Discuţia din capitolul 7 ne arată că e imposibil să formulăm o etică utilitaristă bazată pe maximizarea experienţelor pozitive fără a şti ce entităţi inteligente posedă aşa ceva. După cum am menţionat în capitolul 5, unii oameni preferă ca roboţii să fie inconştienţi pentru a nu se simţi vinovaţi că sunt stăpâni de sclavi. Ei şi-ar putea dori însă contrariul dacă şi-ar uploada minţile pentru a scăpa de limitările biologice: în fond, ce rost are să te uploadezi într-un robot care vorbeşte şi acţionează ca şi cum ar fi doar un zombi inconştient – fiind tu însuţi cel uploadat, nu ţi se pare că lucrurile se schimbă? Nu e oare ca şi cum te-ai sinucide, din perspectiva ta subiectivă, chiar dacă prietenii tăi nu-şi dau seama că experienţa ta subiectivă a murit?

Pentru viitorul cosmic pe termen lung al vieţii (capitolul

6), devine esenţial să înţelegem ce este şi ce nu este conştiinţa: dacă tehnologia permite vieţii inteligente să înflorească în universul nostru vreme de miliarde de ani, cum putem fi siguri că această viaţă e conştientă şi capabilă să-şi dea seama de ce se întâmplă? Altminteri, n-ar fi oare, după cum spunea celebrul fizician Erwin Schrodinger, „o piesă jucată în faţa unor scaune goale, care nu există pentru nimeni, deci, la drept vorbind, nu există deloc”?[2] Cu alte cuvinte, dacă permitem apariţia unor descendenţi de înaltă tehnologie despre care credem în mod greşit că sunt conştienţi, n-ar fi aceasta apocalipsa zombi extremă, transformând măreaţa noastră zestre cosmică într-o irosire astronomică a spaţiului?

CE ESTE CONŞTIINŢA?

Multe dispute privind conştiinţa generează mai curând căldură decât lumină, fiindcă oponenţri înţeleg prin această noţiune lucruri diferite. La fel ca în cazul „vieţii“ şi al „inteligenţei“, nu există o definiţie corectă indiscutabilă a cuvântului „conştiinţă“, iar printre semnificaţiile care i se atribuie se numără capacitatea de a simţi, starea de trezie, conştienţa, accesul la input-ul senzorial şi capacitatea de a contopi informaţiile într-o naraţiune.[3] în explorarea inteligenţei viitoare, vrem să avem o perspectivă cât mai largă, fără a ne limita la tipurile de conştiinţă biologică existente până acum. De aceea definiţia pe care am dat-o în capitolul i, şi cu care am fost consecvent în întreaga carte, e foarte cuprinzătoare:

| conştiinţă = *experienţă subiectivă* |

Cu alte cuvinte, dacă simţi chiar acum că ceva este tu însuţi, atunci eşti conştient. Această definiţie a conştiinţei este cea care ajunge la punctul delicat al întrebărilor despre IA din secţiunea precedentă: Prometeu, AlphaGo sau un automobil Tesla care se conduce singur simt că sunt ceva?

Pentru a vă da seama cât de cuprinzătoare e definiţia noastră, observaţi că nu menţionează comportamentul, percepţia, conştienţa, emoţiile sau atenţia. Conform acestei definiţii, eşti conştient şi atunci când visezi, chiar dacă nu eşti treaz, n-ai acces la input-ul senzorial şi (e de sperat!) nu acţionezi ca un somnambul. în mod asemănător, orice sistem care simte durere e conştient în acest sens, chiar dacă nu se poate mişca. Definiţia noastră lasă deschisă posibilitatea ca şi sisteme IA din viitor să fie conştiente, deşi există doar ca software şi nu sunt conectate la senzori sau la corpuri robotice.

Cu această definiţie, e greu de presupus că problema conştiinţei îţi va rămâne indiferentă. în cartea sa *Homo Dens*[4], Yuval Noah Harari spune: „Dacă un om de ştiinţă susţine că experienţele subiective sunt irelevante, ar trebui să explice de ce tortura sau violul sunt lucruri rele, fără să se refere la vreo experienţă subiectivă.” Fără o asemenea referinţă, n-avem de-a face decât cu o grămadă de particule elementare care se mişcă după legile fizicii – şi ce-i rău în asta?

CARE ESTE PROBLEMA?

Aşadar, ce anume nu înţelegem legat de conştiinţă? Puţini oameni au meditat mai intens asupra acestei întrebări decât David Chalmers, un celebru filozof australian pe care rareori îl vezi fără să aibă pe chip un zâmbet poznaş şi fără să poarte o jachetă neagră de piele – care i-a plăcut atât de mult soţiei mele, încât mi-a dăruit una asemănătoare de Crăciun. Şi-a urmat îndemnul inimii de a studia filozofia, deşi a ajuns în finala Olimpiadei Internaţionale de Matematică – şi deşi în colegiu a avut singura notă mai mică la filozofie. Rămâne netulburat în faţa remarcilor răuvoitoare, şi am fost uimit de capacitatea lui de a asculta

politicos cum ideile sale sunt criticate de persoane neavizate, fără a simţi nevoia să răspundă.

După cum a subliniat David, există două mistere legate de mintea umană. Primul e legat de felul în care creierul procesează informaţia – acestea sunt problemele „uşoare“. De pildă, cum primeşte creierul input-ul senzorial, cum îl interpretează şi cum răspunde la el? Cum îşi relatează prin limbaj starea internă? Deşi aceste întrebări sunt în realitate extrem de dificile, conform defi niţiei noastre nu sunt mistere ale conştiinţei, ci mistere ale inteligenţei: ele vizează felul în care memorează, calculează şi învaţă creierul. Apoi, în prima parte a cărţii de faţă am văzut că, graţie cercetătorilor IA, maşinile au început să facă mari progrese în rezolvarea acestor „probleme uşoare“ – de la jocul de go la conducerea automobilelor, analiza imaginilor şi procesarea limbajului natural.

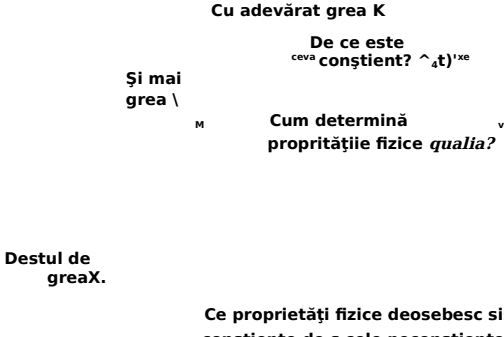
Există însă şi misterul legat de întrebarea „De ce ai o experienţă subiectivă?“, numit de David problema *grea.* în timp ce conduci automobilul, percepi culori, sunete, emoţii şi senzaţia unui sine. Dar de ce ai asemenea experienţe? Un automobil care se conduce singur are vreo experienţă? Dacă te afli într-o cursă cu un automobil care se conduce singur, amândoi primiţi informaţii de la senzori, le procesaţi şi comandaţi motorului. *Experienţa* subiectivă de a conduce e însă ceva logic diferit – este opţională, iar atunci ce anume o provoacă?

Eu abordez această problemă grea a conştiinţei din perspectiva unui fizician. Pentru mine, o persoană conştientă e pur şi simplu hrană rearanjată. Aşadar, de ce un aranjament este conştient, dar nu şi celălalt? Apoi, fizica ne învaţă că hrana e doar un număr mare de cuarci şi electroni aranjaţi într-un anume mod. Prin urmare, care aranjamente de particule sunt conştiente şi care nu sunt conştiente?††††††††††

---

†††††††††††† Un alt punct de vedere este cel al *dualismului substanţei-ideea* că entităţile vii se deosebesc de cele neînsufleţite

Ce îmi place la această perspectivă fizică este că transformă problema grea cu care ne am luptat vreme de milenii într-o versiune mai concentrată care e mai uşor de abordat cu metodele ştiinţei. în loc să începem cu o *problemă* grea, întrebându-ne de ce un aranjament al particulelor se poate simţi conştient, să pornim de la *faptul* că unele aranjamente ale particulelor se simt conştiente, iar altele nu. De pildă, ştii că particulele care alcătuiesc creierul tău se află acum într-un aranjament conştient, dar nu şi atunci când eşti cufundat într-un somn profund, fără vise.



prin faptul că în ele se află o substanţă nefizică („anima“, „elan vital” sau „suflet“). în rândul oamenilor de ştiinţă, sprijinul pentru dualismul substanţei a scăzut treptat. Pentru a înţelege de ce, gândeşte-te că trupul tău e alcătuit din aproximativ io[39] cuarci şi electroni care se mişcă în conformitate cu legi fizice simple. Imaginează-ţi o tehnologie din viitor capabilă să urmărească toate particulele tale: dacă s-ar descoperi că acestea se supun riguros legilor fizicii, atunci presupusul tău suflet n-ar avea nici o influenţă asupra particulelor tale, deci mintea ta conştientă şi capacitatea ei de a-ţi controla mişcările n-ar avea nici o legătură cu vreun suflet. în schimb, dacă s-ar descoperi că particulele tale nu se supun legilor cunoscute ale fizicii pentru că sunt mişcate de sufletul tău, atunci noua entitate care generează aceste forţe ar fi prin definiţie una de natură fizică, pe care o putem studia aşa cum am studiat în trecut noile câmpuri şi particule. *(N. a.)*

Figura 8,1 înţelegerea minţii implică o ierarhie de probleme. Problemele pe care David Chalmers le numeşte „uşoare” pot fi puse fără a menţiona experienţa subiectivă. Faptul că unele sisteme fizice, dar nu toate, sunt conştiente conduce la trei întrebări distincte. Dacă găsim o teorie care să răspundă la întrebarea ce defineşte problema „destul de grea”, atunci aceasta poate fi testată experimental. Dacă teoria funcţionează, atunci, pornind de la ea, putem aborda întrebările mai grele de deasupra.

Această perspectivă fizică ne conduce la trei întrebări dificile privind conştiinţa, după cum se vede în figura 8.1. în primul rând, ce proprietăţi ale aranjamentului particulelor fac diferenţa? Mai precis, ce proprietăţi fizice deosebesc sistemele conştiente de cele neconştiente? Dacă putem răspunde Ia asta, atunci ne putem da seama ce sisteme fizice sunt conştiente. în viitorul imediat, aceasta i-ar putea ajuta şi pe medicii de la camera de gardă să determine ce pacienţi care nu reacţionează sunt conştienţi.

în al doilea rând, cum determină proprietăţile fizice felul în care ne apare experienţa? Mai precis, ce anume determină *qualia,* elementele constitutive ale conştiinţei cum ar fi culoarea roşie a unui trandafir, sunetul unui cinei, gustul unei mandarine sau durerea provocată de o înţepătură?‡‡‡‡‡‡‡‡‡‡

în al treilea rând, de ce este ceva conştient? Cu alte

‡‡‡‡‡‡‡‡‡‡ Folosesc aici cuvântul *qualia* conform definiţiei din dicţionar: cazuri individuale de experienţă subiectivă-experienţa subiectivă în sine, şi nu vreo presupusă substanţă care provoacă experienţa. Atenţie, unii folosesc cuvântul în alt sens! (N. *a.)* [Termenul „qualia“ (în latină: „calităţi”) apare în filozofia minţii şi e preluat ca atare în literatura de specialitate din România – *n.* t]

cuvinte, există o explicaţie profundă încă nedescoperită pentru faptul că grămezi de materie pot fi conştiente, sau e doar un fapt brut inexplicabil legat de felul în care funcţionează lumea?

Informaticianul Scott Aaronson, fost coleg cu mine la MIT, a numit în mod optimist prima întrebare „problema destul de grea“, la fel ca David Chalmers. Să le numim deci pe celelalte două „problema şi mai grea“ şi „problema cu adevărat grea“, după cum se vede în figura 8.1.[§§§§§§§§§§§§]

SE AFLA OARE CONŞTIINŢA DINCOLO DE ŞTIINŢĂ?

Când mi se spune că studiul conştiinţei e pierdere de vreme, principalul argument care mi se aduce este că subiectul e „neştiinţific“ şi aşa va rămâne mereu. E oare adevărat? Karl Popper, influentul filozof britanic de origine austriacă, este autorul celebrei sentinţe „Dacă nu e falsificabil, nu e ştiinţific“,[**********] larg acceptată acum. Cu alte cuvinte, ştiinţa înseamnă pur şi simplu testarea teoriilor în raport cu observaţiile: dacă o teorie nu poate fi testată nici măcar în principiu, atunci e logic imposibil s-o falsifici, ceea ce înseamnă că, după definiţia lui Popper, e neştiinţifică.

Poate aşadar exista o teorie ştiinţifică în stare să

§§§§§§§§§§§§ Iniţial, „problema cu adevărat grea“ am numit-o „problema foarte grea“, însă după ce i-am arătat acest capitol lui David Chalmers, mi a trimis prin e-mail sugestia de a o reboteza „problema cu adevărat grea“, pentru a fi în acord cu semnificaţia pe care i-o dă ei: „Din moment ce primele două probleme (cel puţin astfel formulate) nu ţin cu adevărat de problema grea aşa cum am conceput-o eu, în vreme ce a treia ţine de ea, ai putea-o numi pe a treia «cu adevărat grea», în loc de «foarte grea», pentru a avea un limbaj comun cu mine.“ (N. a.)

********** în accepţiunea lui Popper, a falsifica o teorie înseamnă a dovedi prin experiment că e falsă, adică a găsi cel puţin un caz în care nu se aplică. (N. t)

răspundă la vreuna dintre cele trei întrebări din figura 8. i? Voi încerca să vă conving că răspunsul este un DA! limpede, cel puţin pentru problema destul de grea: „Ce proprietăţi fizice deosebesc sistemele conştiente de cele neconştiente?“ Să presupunem că cineva are o teorie care, pentru orice sistem fizic, răspunde la întrebarea dacă sistemul e conştient prin „da“, „nu“ sau „incert“. Să conectăm creierul tău la un dispozitiv care măsoară unele dintre informaţiile procesate în diferite părţi din creier, şi să introducem aceste informaţii într-un program de calculator care foloseşte teoria conştiinţei pentru a prezice ce părţi din acele informaţii sunt conştiente, şi îţi afişează predicţiile pe un ecran în timp real, ca în figura 8.2. întâi te gândeşti la un măr. Ecranul îţi spune că există în creierul tău informaţii despre un măr, de care eşti conştient, dar că în trunchiul tău cerebral există şi informaţii despre puls, de care nu eşti conştient. Te-ar mira? Deşi primele două predicţii ale teoriei sunt corecte, vrei să faci un test mai riguros. Te gândeşti la mama ta, iar calculatorul îţi spune că există în creierul tău informaţii despre mama ta, dar nu eşti conştient de asta. Teoria face o predicţie incorectă, aşa că e respinsă şi ajunge la coşul de gunoi al istoriei ştiinţei, împreună cu mecanica Iui Aristotel, eterul luminos, cosmologia geocentrică şi nenumărate alte idei greşite. Iată însă elementul esenţial: deşi teoria era falsă, ea era *ştiinţifică*[1]. Dacă n-ar fi fost ştiinţifică, n-ai fi putut-o testa şi respinge.

Unii ar putea critica această concluzie spunând că *ei* n-au nici o dovadă privind lucrul de care eşti conştient, sau chiar privind faptul că ai fi în genere conştient de ceva: deşi te aud spunând că eşti conştient, un zombi neconştient ar vorbi la fel Dar asta nu face să fie neştiinţifică acea teorie despre conştiinţă, fiindcă ei pot trece în locul tău şi testa dacă sunt prezise corect *propriile lor* experienţe conştiente.

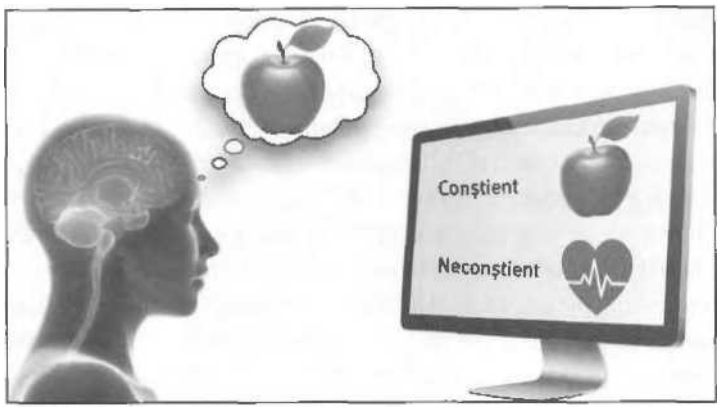


Figura 8.2 Să presupunem că un calculator măsoară informaţia procesată în creierul tău şi prezice de care părţi din ea eşti conştient, conform unei teorii despre conştiinţă. Poţi testa ştiinţific teoria verificând dacă predicţiile ei sunt corecte prin comparaţie cu experienţa ta subiectivă.

Pe de altă parte, dacă teoria refuză să facă vreo predicţie, răspunzând mereu „incert“, ea este netestabilă, deci neştiinţifică. Asta s-ar putea întâmpla fiindcă e aplicabilă doar în anumite situaţii, fiindcă sunt prea greu de efectuat calculele necesare ori fiindcă senzorii cerebrali nu sunt buni. Teoriile ştiinţifice curente se află de regulă undeva la mijloc, oferind răspunsuri testabile la unele întrebări, dar nu la toate. De pildă, teoria noastră centrală din fizică nu răspunde la întrebări despre sisteme care sunt în acelaşi timp extrem de mici (implicând mecanica cuantică) şi extrem de grele (implicând relativitatea generală), fiindcă n-am găsit încă ecuaţiile matematice pentru acest caz. Această teorie centrală nu prezice nici masele exacte ale tuturor atomilor – în acest caz credem că avem ecuaţiile necesare, însă n-am reuşit deocamdată să le calculăm cu precizie soluţiile. Cu cât o teorie trăieşte mai periculos asumân- du-şi riscuri şi făcând predicţii testabile, cu atât e mai utilă şi o luăm mai în serios dacă supravieţuieşte tuturor încercărilor noastre de a o ucide, intr-adevăr, putem testa doar *unele* predicţii ale teoriilor

despre conştiinţă, dar asta e valabil pentru *toate* teoriile din fizică. Prin urmare, în loc să pierdem vremea lamentându-ne pentru ceea ce nu putem testa, n-avem decât să ne punem pe treabă testând ce *putem* testa.

Pe scurt, orice teorie ce prezice care sisteme fizice sunt conştiente (problema destul de grea) este ştiinţifică atâta vreme cât poate prezice care dintre procesele tale cerebrale sunt conştiente. Chestiunea testabilităţii devine însă mai puţin clară pentru întrebările de rang mai înalt în ierarhia din figura 8.1. Ce-ar însemna ca o teorie să prezică felul în care percepi subiectiv culoarea roşie? Iar dacă o teorie pretinde că explică de ce există conştiinţa, cum o testezi experimental? N-ar trebui să evităm asemenea întrebări doar fiindcă sunt dificile – ne vom întoarce la ele, dar cred că, atunci când ne confruntăm cu mai multe întrebări similare rămase fără răspuns, e înţelept să începem prin a o aborda pe cea mai simplă. De aceea cercetările mele de la MIT privind conştiinţa sunt concentrate direct asupra bazei piramidei din figura 8.1. Am discutat recent această strategie cu colegul meu fizician Piet Hut de Ia Princeton, care mi-a spus în glumă că a încerca să construieşti vârful piramidei înainte să construieşti baza e ca şi cum te-ar fi preocupat interpretarea mecanicii cuantice înainte de descoperirea ecuaţiei lui Schrodinger, fundamentul matematic care ne permite să prezicem rezultatele experimentelor noastre.

Când vorbim despre ce se află dincolo de ştiinţă, e important să ţinem cont de faptul că răspunsul depinde de epocă! Cu patru secole în urmă, Galileo Galilei a fost atât de impresionat de teoriile fizice întemeiate pe matematică, încât a spus că natura este „o carte scrisă în limbajul matematicii“. Dacă arunca un strugure şi o alună, putea prezice cu exactitate forma traiectoriilor lor şi momentul când ating solul. Nu avea însă nici un indiciu pentru a explica de ce strugurele e galben, iar aluna maronie, sau

de ce unul e moale, iar cealaltă dură – aceste aspecte ale lumii se aflau dincolo de domeniul ştiinţei la acea vreme. Dar nu pentru totdeauna! Cândjames ClerkMaxwell a descoperit în 1861 ecuaţiile care-i poartă numele, a devenit limpede că lumina şi culorile pot fi înţelese matematic. Ştim acum că ecuaţia lui Schrodinger pomenită mai sus, descoperită în 1925, poate fi folosită pentru a prezice toate proprietăţile materiei, inclusiv ce e moale sau dur. Dacă progresul teoretic ne-a permis să facem tot mai multe predicţii ştiinţifice, progresul tehnologic ne a permis să facem tot mai multe teste experimentale: aproape tot ce studiem acum cu telescoape, microscoape sau acceleratoare de particule s-a aflat odinioară în afara ştiinţei. Cu alte cuvinte, domeniul de competenţă al ştiinţei s-a extins spectaculos începând cu epoca lui Galilei, de la o parte minusculă din ansamblul fenomenelor la un procentaj mare, incluzând particulele subatomice, găurile negre şi originile noastre din urmă cu 13,8 miliarde de ani. Se pune atunci întrebarea: ce rămâne?

După părerea mea, conştiinţa este elefantul din cameră. Nu doar ştii că eşti conştient, dar asta e *tot* ce ştii cu deplină certitudine – restul e deducţie, după cum a arătat Rene' Descartes, contemporanul lui Galilei. Progresul teoretic şi tehnologic va aşeza oare ferm în cele din urmă conştiinţa în domeniul ştiinţei? Nu ştim, aşa cum nici Galilei nu ştia dacă într-o bună zi vom înţelege lumina şi materia.†††††††††††† Un singur lucru e cert: nu vom reuşi dacă

†††††††††††† Dacă realitatea noastră fizică e în întregime de ordin matematic (simplu spus, e bazată pe informaţie), după cum am explorat în cartea mea *Our Mathematical Universe,* atunci nici un aspect al realităţii – şi nici măcar conştiinţa – nu se află în afara domeniului ştiinţei. Problema cu adevărat grea a conştiinţei este, din această perspectivă, exact aceeaşi problemă ca aceea de a înţelege felul în care ceva de natură matematică poate simţi ceea ce e fizic: dacă o parte a unei structuri matematice e conştientă, atunci va percepe restul ca pe lumea fizică externă. *(N.* a.)

nu încercăm! De aceea eu împreună cu alţi oameni de ştiinţă din lumea întreagă ne străduim să formulăm şi să testăm teorii despre conştiinţă.

INDICII EXPERIMENTALE PRIVIND CONŞTIINŢA

Procesarea informaţiei are loc în capetele noastre chiar acum. Cât din aceasta e conştientă, şi cât nu e? înainte de a examina teoriile despre conştiinţă şi ce prezic ele, să vedem ce am aflat până în prezent din experimente, de la observaţii tradiţionale folosind o tehnologie rudimentară sau fără să apeleze deloc la tehnologie, până la măsurători cerebrale de ultimă oră.

Ce comportamente sunt conştiente?

Dacă înmulţeşti mental 32 cu 17, eşti conştient de multe din procesele interioare ale calculului. Dar să presupunem că îţi arăt un portret al lui Albert Einstein şi te invit să-mi spui ce reprezintă. După cum am văzut în capitolul *2,* şi aceasta e o sarcină computaţională: creierul tău evaluează o funcţie al cărei input e informaţia primită de la ochi despre un număr mare de pixeli, iar al cărei output e informaţia trimisă către muşchii care controlează gura ta şi corzile vocale. Informaticienii numesc această sarcină „clasificarea imaginilor“, urmată de „sinteza vorbirii“. Deşi acest calcul e incomparabil mai complicat decât sarcina de a înmulţi două numere, îl efectuezi mult mai repede, aparent fără efort şi fără să fii conştient *de felul în* care îl faci. Experienţa ta subiectivă constă doar în a privi imaginea, a avea senzaţia că recunoşti şi a te auzi pe tine însuţi rostind „Einstein“.

Psihologii ştiu de multă vreme că poţi executa neconştient o gamă largă de alte sarcini şi poţi manifesta o gamă largă de alte comportamente, de la reflexul de a clipi

la respiraţie, întinderea mâinii, apucare şi păstrarea echilibrului. De regulă, eşti conştient de ce ai făcut, dar nu şi de cum ai făcut. Pe de altă parte, comportamentele care implică situaţii neobişnuite, autocontrolul, reguli logice complicate, raţionamentul abstract sau folosirea limbajului sunt în general conştiente. Ele se numesc *corelativele comportamentale ale conştiinţei (behavioral correlates of consciousness),* şi sunt strâns legate de modul lent şi controlat de a gândi, cerând mult efort, cunoscut în psihologie drept „Sistemul 2".[5]

Se ştie de asemenea că multe rutine conştiente pot deveni neconştiente prin exerciţiu intens, cum ar fi mersul, înotul, mersul pe bicicletă, şofatul, dactilografiatul, bărbieritul, legarea şireturilor, jocul pe calculator şi cântatul la pian.[6] într-adevăr, e bine cunoscut faptul că experţii îşi îndeplinesc cel mai bine sarcinile specifice atunci când se află într-o stare „fluidă", conştienţi doar de ce se întâmplă la un nivel înalt şi ignorând detaliile de nivel inferior privind felul în care fac ceea ce fac. De pildă, încearcă să citeşti următoarea frază fiind conştient de fiecare literă în parte, ca şi cum abia înveţi să citeşti. Nu simţi că procesul e mult mai lent decât în cazul în care eşti doar conştient de text la nivelul cuvintelor şi ideilor?

Procesarea neconştientă a informaţiei nu numai că e posibilă, dar pare să fie mai curând regula decât excepţia. Dovezile arată că din cei aproximativ io[7] biţi de informaţie care ajung în creierul nostru în fiecare secundă de la organele senzoriale, nu putem fi conştienţi decât de o parte minusculă, estimările variind de la io la 50 de biţi.[7] Aceasta sugerează că procesarea informaţiei de care suntem conştienţi reprezintă doar vârful aisbergului.

Luate împreună, aceste indicii i-au făcut pe unii cercetători să susţină că procesarea conştientă a informaţiei este un fel de director executiv al minţii noastre, care se ocupă doar de cele mai importante decizii presupunând analiza complexă a datelor din întregul

creier.[8] Asta ar explica de ce, la fel ca directorul executiv al unei companii, de regulă nu vrea să fie distrasă cunoscând tot ce fac subordonaţii – dar poate afla, dacă doreşte. Pentru a percepe această atenţie selectivă în acţiune, priveşte din nou cuvântul „directorul“: fixează-ţi privirea asupra punctului din „i“ şi, fără să ţi mişti ochii, deplasează-ţi atenţia de la punct asupra întregii litere, apoi asupra întregului cuvânt. Deşi informaţia provenită de la retină a rămas aceeaşi, experienţa ta conştientă s-a schimbat. Comparaţia cu directorul executiv explică şi de ce dexteritatea devine neconştientă: după ce s-a chinuit să deprindă cititul şi dactilografierea, directorul executiv delegă aceste sarcini de rutină subordonaţilor neconştienţi, pentru a se putea concentra asupra unor noi provocări de nivel înalt.

Unde se află conştiinţa?

Experimente şi analize ingenioase au sugerat că nu doar la anumite comportamente se limitează conştiinţa, dar şi la anumite părţi din creier. Care sunt principalii suspecţi? Majoritatea primelor indicii

au provenit de la pacienţi cu leziuni cerebrale: vătămări ale creierului localizate, provocate de accidente, atacuri cerebrale, tumori sau infecţii. Rezultatele nu erau însă întotdeauna concludente. De pildă, faptul că leziuni din regiunea dorsală a creierului pot provoca orbirea înseamnă oare că acolo e sediul văzului conştient, sau înseamnă pur şi simplu că informaţia vizuală trece pe-acolo în drum spre ceea ce va deveni conştient, la fel cum trece mai întâi prin ochi?

Deşi leziunile şi intervenţiile medicale n-au localizat cu precizie experienţele conştiente, ele au îngustat sfera posibilităţilor. De exemplu, ştiu că deşi resimt durerea în mână ca şi cum acolo s-ar ivi, senzaţia dureroasă trebuie să apară în alt loc, fiindcă un chirurg mi-a suprimat durerea din mână fără să facă nimic mâinii: mi-a anesteziat nervii din umăr. Mai mult, unii oameni cărora li s-a amputat mâna simt o durere „fantomă" ca şi cum ar apărea în mâna lor inexistentă. Un alt exemplu: am observat odată că, atunci când priveam doar cu ochiul meu drept, lipsea o parte din câmpul meu vizual – un doctor a determinat că retina începea să se dezlipească şi mi-a lipit-o la loc. Dimpotrivă, pacienţi cu anumite leziuni cerebrale suferă de *neglijare hemispaţială:* şi lor le lipseşte informaţia din jumătate din câmpul vizual, dar nu-şi dau seama de asta – de pildă, incapabili să observe restul, mănâncă doar ce se află în jumătatea stângă a farfuriei. E ca şi cum conştiinţa privind jumătate din lumea lor ar fi dispărut. Generează însă oare acele zone vătămate din creier experienţa spaţială, sau doar oferă informaţii sediilor conştiinţei, aşa cum făcea retina mea?

în anii 1930, neurochirurgul canadian de origine americană Wilder Penfield a constatat că pacienţii îi spuneau că diferite părţi ale corpului lor sunt atinse atunci când stimula electric anumite zone din ceea ce se numeşte acum *cortexul somatosenzorial* (figura 8.3).[9] A constatat de asemenea că aceştia îşi mişcau involuntar anumite părţi

ale corpului atunci când stimula zone din ceea ce se numeşte acum *cortexul motor.* înseamnă însă asta oare că procesarea informaţiei în aceste zone cerebrale corespunde atingerii şi mişcării conştiente?

Din fericire, tehnologia modernă ne oferă indicii mult mai amănunţite. Deşi suntem încă departe de a măsura fiecare activare a tuturor celor aproximativ o sută de miliarde de neuroni, tehnologia de citire a creierului progresează rapid, implicând tehnici cu denumiri intimidante precum IRMf, EEG, MEG, ECoG, eFiz şi detecţie pe bază de fluorescenţă. Imagistica prin rezonanţă magnetică funcţională (IRMf) măsoară proprietăţile magnetice ale nucleelor de hidrogen pentru a alcătui o hartă 3-D a creierului în fiecare secundă, la o rezoluţie de ordinul milimetrului. Elec- troencefalografia (EEG) şi magnetoencefalografia (MEG) măsoară câmpul electric şi magnetic din jurul capului pentru a cartografia creierul de mii de ori pe secundă, dar cu o rezoluţie mai slabă, fără a putea distinge detalii mai fine de câţiva centimetri. Dacă eşti temător, te vei bucura să afli că toate aceste trei tehnici sunt ne- invazive. Dacă însă nu te deranjează să-ţi fie deschisă cutia craniană, există şi alte alegeri. Electrocorticografia (ECoG) presupune plasarea a vreo sută de cabluri electrice pe suprafaţa creierului tău, în vreme ce electrofiziologia (eFiz) implică introducerea unor microconductori (adesea mai subţiri decât firul de păr) adânc în creier, pentru a înregistra simultan tensiuni din o sută de locuri. Numeroşi pacienţi cu epilepsie îşi petrec zile întregi în spital în vreme ce ECoG e folosită pentru a descoperi ce parte a creierului declanşează crizele şi trebuie rezecată, şi acceptă în acelaşi timp ca neurologii să facă asupra lor experimente privind conştiinţa, în fine, detecţia pe bază de fluorescenţă presupune manipularea genetică a neuronilor pentru a emite scântei atunci când se activează, ceea ce permite măsurarea activităţii lor cu

ajutorul microscopului. Dintre toate tehnicile, aceasta poate monitoriza rapid cel mai mare număr de neuroni, cel puţin la animale cu creier transparent – cum sunt viermele C. *elegans,* având 302 neuroni, şi larva peştelui-zebră, cu circa 100 000 de neuroni.

Deşi Francis Crick l-a sfătuit pe Christof Koch să nu studieze conştiinţa, Christof n a vrut să renunţe, iar în cele din urmă l-a atras de partea Iui şi pe Francis. în 1990 ei au scris împreună un celebru articol despre ceea ce numeau „corelativele neuronale ale

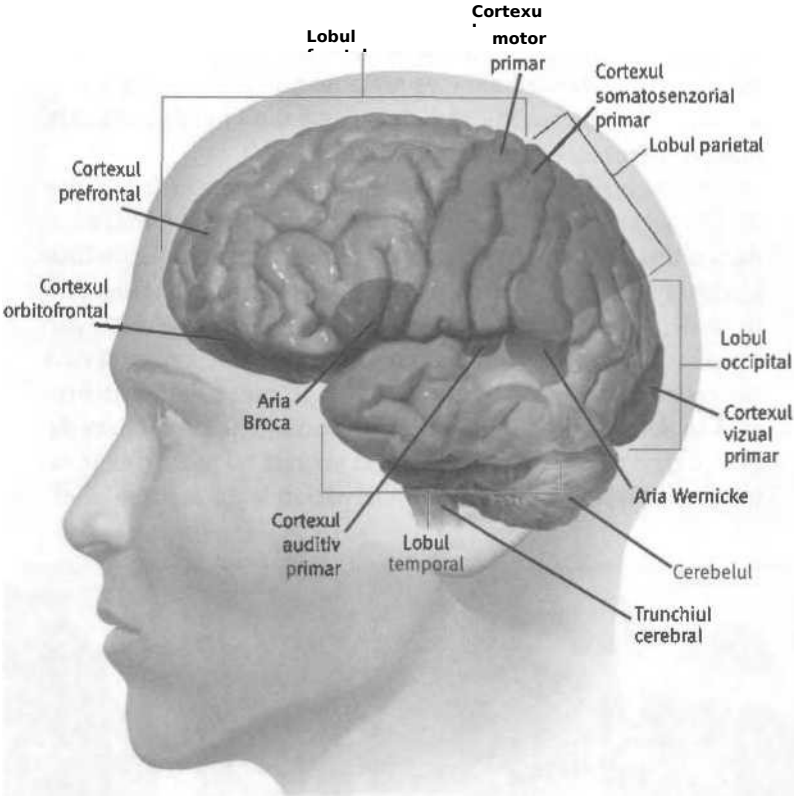


Figura 8.3 Cortexurile vizual, auditiv, somatosenzorial şi motor sunt implicate în văz, auz, simţul tactil şi, respectiv, activarea mişcării – ceea ce nu dovedeşte însă că acolo îşi are sediul *conştiinţa* văzului, auzului, atingerii şi mişcării. Conform unor experimente recente, cortexul vizual primar e complet neconştient, împreună cu cerebelul şi trunchiul cerebral. Imagine obţinută prin amabilitatea agenţiei Lachina (www.lachina.com).

conştiinţei" *(neural correlates of consciousness* – NCC), în care îşi puneau problema să determine care anume procese cerebrale corespund experienţelor conştiente. Vreme de mii de ani, gânditorii avuseseră acces la procesarea informaţiei din creierul lor doar prin experienţa

subiectivă şi comportament. Crick şi Koch au arătat că tehnologia de citire a creierului oferea brusc un acces independent la această informaţie, permiţând studiul ştiinţific al corespondenţei dintre procesarea informaţiei şi experienţa conştientă. Măsurătorile bazate pe noua tehnologie au adus în centrul atenţiei căutarea NCC, mii de articole fiind publicate în cele mai prestigioase reviste de neurologie.[10]

Ce concluzii se pot trage deocamdată? Pentru a ne face o idee despre această muncă detectivistică, sa ne întrebăm mai întâi dacă retina ta e conştientă, sau e doar un sistem zombi care înregistrează informaţia vizuală, o procesează şi o trimite mai departe în creier, unde are loc experienţa vizuală subiectivă. Priveşte partea stângă a figurii 8.4: care pătrat e mai întunecat, cel notat cu A sau cel notat cu B? A, nu-i aşa? Nu, de fapt, sunt colorate identic, ceea ce poţi verifica uitându-te la ele prin micile fante dintre degete. Asta dovedeşte că experienţa ta vizuală nu poate ţine în întregime de retină, fiindcă altminteri pătratele ţi s-ar părea la fel.

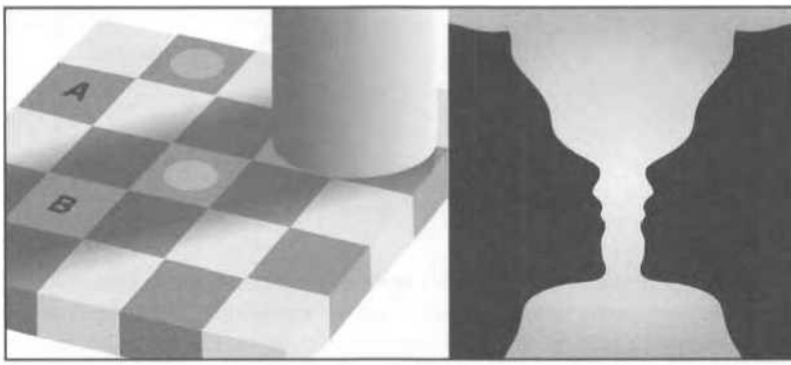


Figura 8.4 Care pătrate maiîntunecat, A sau B? Ce vezi în dreapta, o cupă, două femei sau ambele succesiv? Iluzii optice precum acestea demonstrează că nu în ochi sau în alte prime elemente care aparîn procesul vederii se poate afla conştiinţa ta vizuală, fiindcă ea nu depinde doar de ceea ce este în imagine.

Priveşte acum partea dreaptă a figurii 8.4. Vezi două femei sau o cupă? Dacă te uiţi suficient de mult timp, le vei vedea pe amândouă succesiv, deşi informaţia care ajunge la retina ta rămâne aceeaşi. Măsurând ce se întâmplă în creier în cele două situaţii, putem găsi de unde provine diferenţa – şi nu-i vorba de retină, care se comportă identic în ambele cazuri.

Lovitura fatală dată ipotezei că retina ar fi conştientă vine de la o tehnică numită „suprimarea continuă a instantaneelor" *(continuons flash suppression)* introdusă de Christof Koch, Stanislas Dehaene şi colaboratorii lor: s-a descoperit că dacă unul din ochi urmăreşte un şir complicat de tipare care se modifică rapid, aceasta va distrage sistemul tău vizual în asemenea măsură încât nu vei sesiza deloc o imagine statică prezentată celuilalt ochi.[11] Pe scurt, poţi avea o imagine vizuală în retina ta fără s-o percepi, iar (în timp ce visezi) poţi percepe o imagine fără ca ea să se afle în retina ta. Asta dovedeşte că cele două retine nu găzduiesc conştiinţa ta vizuală, aşa cum nici o cameră video n-o găzduieşte, deşi efectuează calcule complicate implicând peste o sută de milioane de neuroni

Cercetătorii NCC folosesc de asemenea suprimarea continuă a instantaneelor, iluziile optice/auditive instabile şi alte trucuri ca să descopere ce regiuni ale creierului *sunt* responsabile pentru fiecare dintre experienţele conştiente. Strategia de bază e compararea a ceea ce neuronii fac în două situaţii în care în esenţă totul (inclusiv input-ul tău senzorial) este identic – cu excepţia experienţei tale conştiente. Părţile din creier care conform măsurătorilor se comportă diferit sunt apoi identificate drept NCC.

Asemenea cercetări NCC au demonstrat că nimic din conştiinţa ta nu se află în intestine, deşi acolo e localizat sistemul nervos enteric, cu o jumătate de miliard de neuroni care calculează cum să digere optim hrana. Nimic

din conştiinţa ta nu se află în trunchiul cerebral, partea inferioară a creierului care se leagă de măduva spinării şi controlează respiraţia, ritmul cardiac şi tensiunea arterială. încă mai surprinzător, conştiinţa ta nu pare să se extindă Ia cerebel (figura 8.3), care conţine circa două treimi din toţi neuronii tăi: pacienţii al căror cerebel e distrus vorbesc nedesluşit şi se mişcă stângaci, semănând cu beţivii, dar sunt pe deplin conştienţi.

Care părţi ale creierului *sunt* responsabile pentru conştiinţă rămâne o întrebare fără un răspuns cert. Unele cercetări NCC recente sugerează să sediul conştiinţei ar fi într-o „zonă fierbinte" implicând talamusul (aproape de centrul creierului) şi partea dorsală a cortexului (învelişul extern al creierului constând într-o suprafaţă încreţită alcătuită din şase straturi, care, dacă ar fi întinsă, ar avea aria unei feţe de masă).[12] Aceleaşi cercetări sugerează că o excepţie o constituie cortexul vizual primar, situat în extremitatea dorsală a creierului, care e la fel de neconştient ca globii oculari sau retinele.

Când apare conştiinţa?

Am prezentat până acum indicii privind acele tipuri de procesare a informaţiei care sunt conştiente şi acele locuri unde apare conştiinţa. Dar *când* apare conştiinţa? în copilărie credeam că devenim conştienţi de evenimente pe măsură ce ele se petrec, fără vreun decalaj temporal. Deşi am în continuare aceeaşi senzaţie subiectivă, e limpede că nu poate fi corectă, din moment ce durează un timp până ce creierul meu procesează informaţia care intră prin organele mele de simţ. Cercetătorii NCC au măsurat acest decalaj, iar Christof Koch susţine că trece aproximativ un sfert de secundă între momentul în care lumina intră în ochii tăi provenind de la un obiect complex şi momentul în care percepi în mod conştient că îl vezi.[13] Asta înseamnă că atunci când şofezi pe o autostradă cu viteza de nouăzeci de

kilometri pe oră şi vezi brusc o veveriţă la câţiva metri în faţa ta, ai călcat-o deja!

Pe scurt, conştiinţa ta trăieşte în trecut, decalajul faţă de lumea exterioară fiind, după estimarea lui Christof Koch, de aproximativ un sfert de secundă. Şi totuşi, poţi adesea reacţiona la lucruri mai repede decât poţi deveni conştient de ele, ceea ce dovedeşte că procesarea informaţiei responsabilă pentru cele mai rapide reacţii trebuie să fie neconştientă. De pildă, dacă un obiect străin se apropie de ochii tăi, reflexul de a clipi poate închide pleoapa ta într-o zecime de secundă. Este ca şi cum unul dintre sistemele tale cerebrale primeşte informaţiile ameninţătoare de la sistemul vizual, calculează că ochiul tău riscă să fie lovit, transmite muşchilor ochiului instrucţiuni să clipească şi, simultan, transmite părţii conştiente din creier mesajul „Hei, o să clipim“. în momentul în care acest mesaj e citit şi inclus în experienţa ta conştientă, ai clipit deja.

într-adevăr, sistemul care citeşte acel mesaj e continuu bombardat cu mesaje din întreg corpul tău, unele mai întârziate decât altele. Durează mai mult ca semnalele nervoase să ajungă la creier de la degete decât de Ia obraz din cauza distanţei, şi durează mai mult să analizezi imaginile decât sunetele, fiindcă e mai complicat – de aceea la cursele atletice startul se dă printr-o împuşcătură, nu printr-un indiciu vizual. Şi totuşi, atunci când îţi atingi nasul, resimţi în mod conştient senzaţia nasului simultan cu senzaţia vârfului degetului, iar atunci când baţi din palme, vezi, auzi şi simţi bătaia exact în acelaşi timp.[14] Asta înseamnă că întreaga ta experienţă conştientă privind un eveniment e creată abia după ce au ajuns şi cele mai întârziate mesaje şi au fost analizate.

Un faimos set de experimente NCC iniţiate de fiziologul Benjamin Liber demonstrează că acele acţiuni pe care le poţi efectua neconştient nu se limitează Ia reacţii rapide cum sunt clipitul şi loviturile din ping pong, ci includ şi

anumite decizii pe care le-ai pune pe seama liberului-arbitru – măsurătorile cerebrale pot uneori prezice hotărârea ta înainte să devii conştient că ai luat-0.[15]

TEORII DESPRE CONŞTIINŢĂ

Am văzut că, deşi încă nu înţelegem conştiinţa, avem mari cantităţi de date experimentale privind diverse aspecte ale ei. Toate aceste date provin însă de la *creier,* prin urmare ce ne-ar putea spune despre conştiinţa din *maşini?* Aceasta presupune o extrapolare majoră, dincolo de domeniul nostru experimental curent. Cu alte cuvinte, e nevoie de o teorie.

De ce o teorie?

Pentru a înţelege de ce, să comparăm teoriile despre conştiinţă cu cele despre gravitaţie. Oamenii de ştiinţă au început sa ia în serios teoria newtoniană a gravitaţiei fiindcă au obţinut de la ea mai mult decât au introdus în ea: ecuaţii simple care încăpeau pe un şerveţel de hârtie puteau prezice corect rezultatul fiecărui experiment privind gravitaţia efectuat vreodată. Prin urmare, ei au luat în serios predicţiile ei şi mult in afara domeniului în care fusese testată, iar aceste extrapolări îndrăzneţe s-au dovedit a funcţiona chiar şi pentru mişcările galaxiilor din roiuri de galaxii întinse pe milioane de ani-lumină. Predicţiile erau însă uşor incorecte pentru mişcarea lui Mercur în jurul Soarelui. Oamenii de ştiinţă au început să ia în serios mai buna teorie einsteiniană a gravitaţiei, relativitatea generală, fiindcă era mai elegantă şi mai economică, şi făcea predicţii corecte şi acolo unde teoria lui Newton dădea greş. Prin urmare, au luat în serios şi predicţiile ei mult dincolo de domeniul în care fusese testată, pentru fenomene exotice cum sunt găurile negre,

undele gravitaţionale în însăşi ţesătura spaţiu- luitimp şi expansiunea universului nostru dintr-o origine fierbinte – toate acestea fiind ulterior confirmate experimental.

în mod asemănător, dacă o teorie matematică a conştiinţei ale cărei ecuaţii încap pe un şerveţel de hârtie ar prezice cu succes rezultatele tuturor experimentelor efectuate pe creier, am începe să luăm în serios nu doar teoria însăşi, dar şi predicţiile ei pentru o conştiinţă aflată în afara creierului – de pildă, într-o maşină.

Conştiinţa din perspectiva fizicii

Deşi unele teorii despre conştiinţă apar încă din Antichitate, majoritatea celor modeme sunt întemeiate pe neuropsihologie şi neuroştiinţe, încercând să explice şi să prezică conştiinţa în termenii evenimentelor neuronale care au loc în creier.[16] Cu toate că aceste teorii au făcut unele predicţii corecte privind corelativele neuronale ale conştiinţei, ele nu pot şi nu urmăresc să facă predicţii privind conştiinţa maşinilor. Pentru saltul de la creiere la maşini, trebuie să generalizăm de Ia NCC la PCC: corelativele fizice ale conştiinţei *(physical correlates ofconsciousness),* definite ca tipare conştiente ale particulelor în mişcare. Dacă o teorie poate prezice corect ce este şi ce nu este conştiinţa referindu-se doar la constituenţi fizici de bază cum sunt particulele elementare şi câmpurile de forţe, atunci ea poate face predicţii nu numai pentru creiere, dar şi pentru orice alte aranjamente ale materiei, inclusiv viitoarele sisteme IA. Să considerăm deci o perspectivă a fizicii: ce aranjamente de particule sunt conştiente?

Ajungem astfel Ia o altă întrebare: cum poate ceva atât de com plex precum conştiinţa să fie alcătuit din ceva atât de simplu precum particulele? Este vorba, cred, despre un fenomen care are proprietăţi aflate deasupra şi dincolo de cele ale particulelor. în fizică, asemenea fenomene se

numesc „emergente“.[17] Pentru a le înţelege, vă propun să analizăm un fenomen emergent mai simplu decât conştiinţa: umiditatea.

O picătură de apă e umedă, însă un cristal de gheaţă şi un nor de aburi nu sunt, deşi au în compoziţie molecule de apă identice. De ce? Fiindcă proprietatea de a fi umed depinde doar de aranjamentul moleculelor. N-are sens să spui că o singură moleculă de apă e umedă, pentru că fenomenul umidităţii apare numai dacă există multe molecule, aranjate în tiparul pe care îl numim lichid. Prin urmare, solidele, lichidele şi gazele sunt, toate, fenomene emergente: ele reprezintă mai mult decât suma componentelor, fiindcă au proprietăţi aflate deasupra şi dincolo de proprietăţile componentelor lor. Posedă proprietăţi pe care componentele lor nu le au.

La fel ca în cazul solidelor, lichidelor şi gazelor, conştiinţa cred că e un fenomen emergent, cu proprietăţi aflate deasupra şi dincolo de cele ale particulelor care o alcătuiesc. De pildă, intrarea într-un somn profund abrogă conştiinţa pur şi simplu prin rearan- jarea particulelor. La fel, conştiinţa mea ar dispărea dacă aş muri îngheţat, ceea ce mi-ar rearanja particulele într-un mod mult mai nefericit.

Când pui multe particule laolaltă pentru a alcătui orice, de la apă la creier, apar noi fenomene cu proprietăţi observabile. Nouă, fizicienilor, ne place să studiem aceste proprietăţi emergente, care pot fi adesea identificate printr-un mic set de numere măsurabile – mărimi precum vâscozitatea unei substanţe, compresibilitatea ei etc. De pildă, dacă o substanţă e atât de vâscoasă încât e rigida, o numim solidă, altminteri o numim fluidă. Dacă un fluid e incom- presibil, îl numim lichid, altminteri îl numim gaz sau plasmă, în funcţie de conductivitatea lui electrică.

Conştiinţa ca informaţie

Pot exista aşadar mărimi analoge care să cuantifice conştiinţa? Specialistul italian în neuroştiinţe Giulio Tononi a propus o asemenea mărime, pe care o numeşte *informaţie integrată,* notată cu litera grecească O (Phi) şi care măsoară în esenţă cât de mult ştiu una despre alta diferite părţi ale unui sistem (vezi figura 8.5).

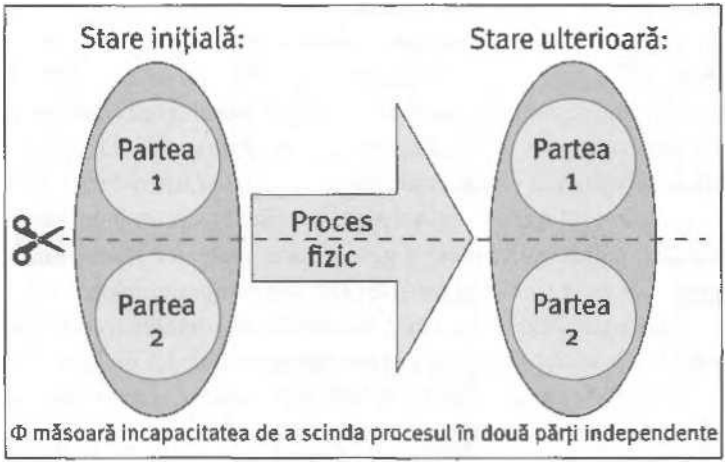


Figura 8.5 Pentru un proces fizic care, odată cu trecerea timpului, transformă starea inițială a unui sistem într-o nouă stare, *informația integrată* <P măsoară incapacitatea de a scinda procesulîn două părți independente. Dacă starea viitoare a fiecărei părți depinde doar de propriul ei trecut, şi nu de ceea ce a făcut cealaltă parte, atunci – o: ceea ce numeam un sistem reprezintă de fapt două sisteme independente care nu comunică unul cu altul.

L-am cunoscut pe Giulio în 2014 la o conferinţă de fizică din Puerto Rico, unde i-am invitat pe el şi pe Christof Koch, şi mi a părut a fi renascentistul perfect care ar fi putut sta alături de Galilei şi Leonardo da Vinci. Modestia sa nu putea ascunde incredibilele Iui cunoştinţe artistice, literare şi filozofice, iar reputaţia sa culinară mi-a fost adeverită de un jurnalist TV care mi-a spus recent că în doar câteva minute Giulio i-a pregătit cea mai delicioasă salată pe care a mâncat-o vreodată. Mi-am dat curând seama că în spatele exprimării sale cumpănite se ascunde un intelect

cutezător, gata să urmeze dovezile oriunde l-ar aduce, indiferent faţă de ideile preconcepute şi de tabuurile dominante. La fel cum Galilei şi-a construit teoria matematică a mişcării în ciuda interdicţiei de a pune în discuţie geocentrismul, Giulio a elaborat cea mai precisă matematic dintre teoriile despre conştiinţă de care dispunem, *teoria informaţiei integrate (integrated information theory-IYT).*

Am susţinut vreme de decenii că de fapt conştiinţa este ceea ce simte informaţia atunci când e procesată în anumite moduri complexe.[18] IIT îmi dă dreptate şi înlocuieşte exprimarea mea vagă „anumite moduri complexe“ cu o definiţie precisă: procesarea informaţiei trebuie să fie integrată, adică trebuie să fie mare. Argumentul lui Giulio este pe cât de puternic, pe atât de simplu: sistemul conştient trebuie să fie integrat într-un ansamblu unificat, fiindcă dacă ar fi constituit din două părţi independente, s-ar simţi ca două entităţi conştiente separate, nu ca una singură. Cu alte cuvinte, dacă o parte conştientă a unui creier sau a unui calculator nu poate comunica cu restul, atunci restul nu poate fi o parte a experienţei subiective a acesteia.

Giulio şi colaboratorii săi au măsurat o versiune simplificată a lui <t> folosind EEG pentru a înregistra răspunsul creierului la stimularea magnetică- Acest „detector de conştiinţă“ funcţionează foarte bine a determinat că pacienţii erau conştienţi când erau treji sau visau, dar neconştienţi când erau anesteziaţi sau în somn profund. A detectat conştiinţa chiar şi la doi pacienţi suferind de sindromul *locked-in*‡‡‡‡‡‡‡‡‡‡ care nu se puteau mişca sau comunica normal.[19] Avem aşadar de a face cu o

‡‡‡‡‡‡‡‡‡‡‡ Literal, „sindromul închiderii“, numit şi pseudocomă, în care aproape toţi muşchii de control voluntar sunt paralizaţi, iar pacientul, care de altfel e perfect conştient, nu poate comunica decât clipind sau eventual mişcân- du-şi globii oculari. (N. t)

tehnologie promiţătoare pentru ca în viitor medicii să-şi dea seama dacă anumiţi pacienţi sunt conştienţi sau nu.

Ancorarea conştiinţei în fizică

irr e definită doar pentru sisteme discrete care se pot afla într-un număr finit de stări, ca biţii din memoria unui calculator sau neuronii ultrasimplificaţi care sunt fie activaţi, fie dezactivaţi. Ca atare, din păcate IIT nu e definită pentru majoritatea sistemelor fizice tradiţionale, care pot varia în mod continuu – de pildă, poziţia unei particule sau intensitatea câmpului magnetic pot lua orice valoare dintr-un spectru infinit.[10] Dacă încerci să aplici formula IIT la asemenea sisteme, obţii de regulă rezultatul absurd că O e infinit. Sistemele cuantice pot fi discrete, dar IIT iniţială nu e definită pentru sisteme cuantice. Cum putem deci ancora IIT şi alte teorii despre conştiinţă bazate pe informaţie pe un fundament fizic solid?

O putem face pornind de la ceea ce am aflat în capitolul 2 despre felul în care grămezi de materie pot avea proprietăţi emergente care sunt asociate cu informaţia. Am văzut că pentru ca ceva să fie folosit ca dispozitiv de memorie care stochează informaţia, trebuie să aibă mai multe stări cu timp de viaţă lung. Am văzut de asemenea că pentru ca o substanţă să fie *computroniu* (să poată efectua calcule), este nevoie în plus de o dinamică complexă: legile fizicii trebuie s-o poată transforma în moduri suficient de complicate pentru a fi efectuată orice procesare a informaţiei. în fine, am văzut că o reţea neuronală, de exemplu, este un puternic substrat pentru învăţare, fiindcă, supunându-se pur şi simplu legilor fizicii, se poate rea- ranja pentru a deveni tot mai bine adaptată efectuării calculelor dorite. Se pune acum încă o întrebare; ce face ca o bucată de materie să aibă experienţe subiective? Cu alte cuvinte, în ce condiţii o bucată de materie va putea face următoarele patru lucruri?

1. memorează
2. calculează
3. învaţă
4. are experienţe

Despre primele trei am vorbit în capitolul *2,* să ne ocupăm acum de al patrulea lucru. La fel cum Margolus şi Toffoli au creat termenul *computroniu* pentru o substanţă care poate efectua calcule arbitrare, îmi place să folosesc termenul *sentroniu* pentru cea mai generală substanţă care are experienţe subiective (simte).[§§§§§§§§§§§§]

Dar cum de este resimţită conştiinţa ca un fenomen non-fizic, dacă de fapt e un fenomen fizic? Cum se poate simţi independentă de substratul ei fizic? Eu cred că *este* independentă de sub stratul fizic, materia din care ea alcătuieşte un tipar! Am întâlnit multe exemple frumoase de structuri independente de substrat în capitolul 2, între care undele, memoriile şi calculele. Am văzut că acestea nu erau doar mai mult decât părţile lor componente (adică erau emergente), dar şi independente de părţile lor, având propria lor viaţă. De pildă, am văzut că o minte simulată sau un personaj dintr-un joc pe computer nu şi-ar putea da seama dacă sunt rulate pe Windows, Mac OS, pe un telefon Android sau pe vreun alt sistem de operare, fiindcă ar fi independente de substrat. N-ar putea şti nici dacă porţile logice ale calculatorului sunt alcătuite din tranzistori, circuite optice sau alt hardware. Şi nici care sunt legile fundamentale ale fizicii – acestea pot fi oricum, cu condiţia să permită construirea calculatoarelor universale.

Pe scurt, conştiinţa cred că e un fenomen fizic care se simte non-fizic deoarece este la fel ca undele şi calculele:

---

§§§§§§§§§§§§ Deşi am folosit în trecut „perceptroniu” ca sinonim pentru sentroniu, această denumire sugerează o definiţie prea îngustă, fiindcă percepţiile sunt doar acele experienţe subiective pe care le resimţim pe baza input-ului senzorial - excluzând, de pildă, visele şi gândurile generate intern. *(N. a.)*

are proprietăţi independente de substratul ei fizic. Asta decurge logic din ideea de conştiinţă-ca-informaţie. Ajungem aşadar la o idee radicală, care îmi place foarte mult: dacă conştiinţa este ceea ce simte informaţia când e procesată în anumite moduri, atunci trebuie sa fie independentă de substrat; contează doar structura procesării informaţiei, nu şi structura materiei care efectuează procesarea informaţiei Cu alte cuvinte, conştiinţa este de două ori independentă de substrat!

După cum am văzut, fizica descrie tipare în spaţiu-timp corespunzând particulelor ce se mişcă de jur-împrejur. Dacă aranjamentele particulelor se supun anumitor principii, ele dau naştere fenomenelor emergente care sunt în mare măsură independente de substratul particulelor şi le resimt pe acestea într-unmod total diferit. Un exemplu este procesarea informaţiei în computroniu. Am dus însă acum această idee la un alt nivel: *dacă însăşi procesarea informaţiei se supune anumitor principii, eapoateda naştere unui fenomen emergent de nivel mai înalt pe care îl numim conştiinţă.* Asta aşază experienţa ta conştientă nu cu un nivel, ci cu două niveluri deasupra materiei. Nu-i de mirare că mintea ta se simte non-fîzică!

Se pune deci următoarea întrebare: care sunt principiile cărora trebuie sa li se supună procesarea informaţiei pentru a fi conştientă? Nu pretind că ştiu ce condiţii sunt *suficiente* pentru a garanta conştiinţa, dar iată patru condiţii *necesare pe* care aş paria şi pe care le-am explorat în cercetările mele:

Principiu	**Definiţie**
Principiul informaţiei	Un sistem conştient are 0 mare capacitate de stocare a informaţiei.
Principiul dinamicii	Un sistem conştient are 0 mare capacitate de procesare a informaţiei.
Principiul independenţei	Un sistem conştient are 0 mare independenţă faţă de restul lumii.

Principiul integrării	Un sistem conştient nu poate fi alcătuit din părţi aproape independente.

După cum am spus, conştiinţa cred ca este ceea ce simte informaţia atunci când e procesată în anumite moduri. Pentru a fi aşadar conştient, un sistem fizic trebuie să poată stoca şi procesa informaţie – de aici rezultă primele două principii. Observaţi că memoria nu e necesar sa fie de lungă durată: vă recomand să priviţi filmuleţul cu Clive Wearing, care pare perfect conştient, deşi amintirile lui durează mai puţin de un minut.[21] Cred că un sistem conştient trebuie de asemenea să fie relativ independent de restul lumii, fiindcă altminteri n-ar simţi în mod subiectiv că are o existenţă independentă. în fine, cred că un sistem conştient trebuie să fie integrat intr-un ansamblu unificat, după cum susţinea Giulio Tononi, fiindcă dacă ar fi constituit din două părţi independente, s-ar simţi ca două entităţi conştiente separate, nu ca una singură. Primele trei principii implică *autonomia:* sistemul e capabil să reţină şi să proceseze informaţie fără să interfereze prea mult cu exteriorul, determinându-şi astfel propriul viitor. Luate împreună, cele patru principii arată că sistemul e autonom, dar nu şi părţile lui.

Dacă aceste patru principii sunt corecte, lucrurile se limpezesc: trebuie să căutăm teorii riguroase matematic care să le încorporeze, şi să testăm experimental teoriile. Trebuie de asemenea să vedem dacă e nevoie şi de alte principii. Indiferent dacă IIT este corectă sau nu, cercetătorii ar trebui să elaboreze teorii concurente şi să testeze toate teoriile disponibile prin experimente mai bune.

COTROVERSE PRIVIND CONŞTIINŢA

Am vorbit deja despre eterna dispută în jurul întrebării

dacă cercetările legate de conştiinţă sunt un nonsens neştiinţific şi o pierdere de vreme. Recent au apărut controverse în cercetările de vârf privind conştiinţa – vă propun să le examinăm pe cele mai interesante dintre acestea.

Teoria informaţiei integrate a lui Tononi a fost întâmpinată nu doar cu laude, dar şi cu critici, unele chiar nimicitoare. Iată ce scria nu de mult Scott Aaronson pe blogul său: „După părerea mea, faptul că teoria informaţiei integrate e falsă – demonstrabil falsă, din motive care se află în miezul ei – o aşază în ceva de genul top *2%* dintre toate teoriile matematice despre conştiinţă propuse vreodată. Cred că aproape toate teoriile concurente despre conştiinţă au fost atât de vagi, de superficiale şi de maleabile, încât pot doar aspira la falsitate.“[22] Spre cinstea lor, Scott şi Giulio s-au purtat absolut civilizat în dezbaterea despre IIT de la Universitatea din New York, fiecare ascultând politicos argumentele celuilalt. Aaronson a arătat că în anumite reţele simple de porţi logice informaţia e puternic integrată (<t> mare), iar din moment ce în mod evident acestea nu ar fi conştiente, IIT e falsă. Giulio a ripostat afirmând că, dacă aceste reţele ar fi construite, ele *ar fi* conştiente, iar presupunerea contrară a lui Scott era marcată de antropocentrism, cam Ia fel ca afirmaţia unui proprietar de abator cum că animalele nu pot fi conştiente fiindcă nu vorbesc şi sunt foarte diferite de oameni. în opinia mea, cu care ambii au fost de acord, dezacordul lor ţinea de faptul că pentru Scott integrarea era doar o condiţie *necesară,* în vreme ce pentru Giulio era şi o condiţie *suficientă.* Această din urmă afirmaţie e mai tranşantă, şi sper ca în curând să poată fi testată experimentai.[2]*

Altă afirmaţie IIT controversată este că arhitecturile calculatoarelor de azi nu pot fi conştiente, fiindcă felul în care se conectează porţile lor logice oferă o integrare foarte scăzută?[4] Cu alte cuvinte, dacă te uploadezi într-un

robot viitor de mare putere care simulează cu precizie fiecare dintre neuronii şi sinapsele tale, atunci, chiar dacă această clonă digitală arată, vorbeşte şi acţionează într-un fel prin care să nu poată fi deosebită de tine, Giulio susţine că va fi un zombi neconştient fără experienţe subiective – o dezamăgire pentru tine, dacă te-ai uploadat în căutarea nemuririi subiective* Această afirmaţie a fost contrazisă de David Chalmers şi de profesorul de IA Murray Shanahan, ei imaginându-şi ce s-ar întâmpla dacă ţi-ar fi treptat înlocuite circuitele neuronale din creier cu un ipotetic hardware digital care le simulează perfect?[5] Deşi comportamentul tău n-ar fi afectat de înlocuire, din moment ce simularea e presupusă a fi perfectă, *experienţa* ta ar trebui să se modifice – din conştienţă, la început, în neconştientă, la sfârşit –, dacă

♦ Există o tensiune potenţială între această afirmaţie şi ideea de independenţă de substrat a conştiinţei, din moment ce, deşi procesarea informaţiei poate fi diferită la nivelul cel mai de jos, ea este prin definiţie identică la nivelurile superioare, acolo unde determină comportamentul. (N. a.)

ar fi să-i dăm crezare lui Giulio. Dar ce ai simţi în etapele intermediare, pe măsură ce tot mai multe circuite neuronale simt înlocuite? Când părţile din creier responsabile de experienţa conştientă a jumătăţii superioare a câmpului tău vizual ar fi înlocuite, ai observa că jumătate din peisajul vizual a dispărut brusc, dar că în chip misterios ştii totuşi ce se află acolo, aşa cum se întâmplă cu pacienţii suferind de „vedere oarbă“**********?[26] Ar fi într-adevăr tulburător, pentru că, dacă ai simţi în mod conştient vreo diferenţă, atunci le-ai putea vorbi despre asta şi prietenilor tăi – însă am presupus că nu apare nici o modificare în comportamentul tău. Singura posibilitate logică, în acord cu acele presupuneri, este că, exact în momentul când oricare

********** Capacitatea pacienţilor cu leziuni corticak de a reacţiona la stimuli vizuali pe care nu-i văd în mod conştient. (N. t)

dintre lucruri dispare din conştiinţa ta, mintea ta e alterată în chip misterios, aşa încât fie minţi şi negi că experienţa ta s-a modificat, fie uiţi că lucrurile fuseseră diferite.

Pe de altă parte, Murray Shanahan recunoaşte că aceeaşi critică legată de înlocuirea treptată poate fi aplicată *oricărei* teorii care afirmă că poţi acţiona conştient fără să fii conştient, aşa încât ai putea conchide că a acţiona conştient şi a fi conştient sunt unul şi acelaşi lucru, iar comportamentul observabil din exterior este deci tot ce contează. Atunci însă cazi în capcana de a prezice că eşti neconştient în timp ce visezi, deşi ştii că nu-i aşa.

A treia controversă privind IIT se leagă de întrebarea dacă o entitate conştientă poate fi alcătuită din părţi care sunt conştiente în mod separat. De pildă, poate societatea în ansamblul ei să capete conştiinţă fără ca oamenii să şi-o piardă pe a lor? Poate un creier conştient să fie alcătuit din părţi care să aibă şi ele propria lor conştiinţă? Conform IIT, răspunsul este „nu“, dar nu toată lumea e convinsă. De pildă, unii pacienţi cu leziuni grave care le reduc comunicarea dintre cele două emisfere cerebrale resimt „sindromul mâinii străine”, în care emisfera dreaptă comandă mâinii stângi să facă lucruri pe care pacienţii afirmă că nu le provoacă sau nu le înţeleg – uneori ajung chiar să-şi folosească cealaltă mână pentru a-şi opri mâna „străină“. Cum putem fi siguri că nu există două conştiinţe separate în capul lor, una în emisfera dreaptă, incapabilă să vorbească, cealaltă în emisfera stângă, responsabilă de întreaga comunicare verbală şi afirmând că vorbeşte în numele ambelor emisfere? Imaginaţi-vă că prin tehnologia viitorului s-ar realiza comunicarea directă între două creiere umane, iar treptat ar creşte capacitatea acestei legături până când comunicarea ar fi la fel de eficientă între creiere ca în interiorul lor. Este oare posibil ca Ia un moment dat cele două conştiinţe individuale să dispară brusc şi să fie înlocuite de una singură, unificată, aşa cum prezice IIT, sau tranziţia ar fi treptată, aşa încât

conştiinţele individuale să coexiste sub o anumită formă chiar şi atunci când ar începe să apară experienţa unificată?

Altă controversă fascinantă se leagă de întrebarea dacă nu cumva experimentele subestimează totalitatea lucrurilor de care suntem conştienţi. Am văzut mai devreme că deşi *simţim* că suntem vizual conştienţi de o imensă cantitate de informaţii despre culori, forme, obiecte şi practic tot ce se află în faţa noastră, experimentele au arătat că putem ţine minte şi evoca doar o infimă parte din acestea?[7] Unii cercetători au încercat să lămurească această discrepanţă întrebându-se dacă nu cumva uneori avem „conştiinţă fără acces“, adică experienţa subiectivă a lucrurilor care sunt prea complexe pentru a intra în memoria noastră de lucru spre a fi utilizate mai târziu.[18] De pildă, atunci când ajungi la *orbire din neatenţie,* fiind prea distras ca să observi un obiect aflat chiar în faţa ta, asta nu înseamnă că n-ai experienţa vizuală a lui, ci doar că ea n-a fost stocată în memoria ta de lucru?[9] Nu-i vorba aici mai curând de uitare decât de orbire? Alţi cercetători resping ideea că nu poţi avea încredere în ce spun oamenii că resimt, şi ne avertizează în privinţa implicaţiilor ei. Murray Shanahan îşi imaginează un test clinic în care pacienţii mărturisesc că au scăpat complet de durere printr-un nou medicament miraculos, care e totuşi respins de o comisie guvernamentală: „Pacienţii doar îşi închipuie că au scăpat de durere. Graţie neuroştiinţelor, noi ştim mai bine.”[30] Pe de altă parte, au existat cazuri în care pacienţilor care s-au trezit din întâmplare în timpul operaţiei li s-a administrat un medicament ca să uite calvarul prin care au trecut. Ar trebui să ne încredem în declaraţiile lor ulterioare cum că n-au simţit nici o durere?’[1]

Dacă un sistem IA din viitor va fi conştient, care va fi experienţa lui subiectivă? Aceasta e esenţa problemei „şi mai grea“ a conştiinţei, iar astfel ajungem la al doilea nivel de dificultate din figura 8.1. Nu numai că deocamdată ne lipseşte o teorie care să răspundă la această întrebare, dar nici măcar nu suntem siguri că e logic posibil să răspunzi pe deplin la ea. În fond, cum ar suna un răspuns satisfăcător? Cum să-i explici unui om născut orb cum arată culoarea roşie?

Din fericire, incapacitatea noastră actuală de a da un răspuns complet nu ne împiedică să dăm răspunsuri parţiale. Extrate- reştri inteligenţi care ar studia sistemul senzorial uman ar deduce probabil despre culori că sunt qualia percepute a fi asociate cu fiecare punct de pe o suprafaţă bidimensională (câmpul nostru vizual), în vreme ce sunetele nu sunt percepute a fi localizate spaţial, iar durerile sunt percepute a fi asociate cu diferite părţi ale corpului nostru. Din descoperirea faptului că retinele noastre au trei tipuri de celule conice fotosensibile, ar putea deduce că percepem trei culori primare, iar toate celelalte qualia de culoare rezultă din combinarea acestora. Măsurând cât timp durează ca neuronii să transmită informaţia prin creier, ar putea trage concluzia că avem aproximativ zece gânduri sau percepţii conştiente pe secundă, iar atunci când privim la televizor filme cu 24 de cadre pe secundă, nu le percepem ca pe un şir de imagini statice, ci ca pe o mişcare continuă. Măsurând cât de rapid e eliberată adrenalina în sânge şi cât timp trece până dispare, ar putea deduce că accesele noastre de furie se declanşează în câteva secunde şi durează câteva minute.

Aplicând raţionamente similare bazate pe fizică, putem lansa unele ipoteze privind felul în care ar putea simţi o conştiinţă artificială. în primul rând, spaţiul experienţelor

IA posibile e *uriaş* în comparaţie cu ceea ce putem resimţi noi, oamenii. Avem câte o clasă de qualia pentru fiecare dintre simţurile noastre, dar o IA poate avea incomparabil mai multe tipuri de senzori şi de repre zentări interne ale informaţiei, aşa încât trebuie să evităm capcana de a presupune că o IA s ar simţi la fel ca un om.

în al doilea rând, o conştiinţă artificială de dimensiunea creierului nostru ar avea de milioane de ori mai multe experienţe pe secundă decât noi, fiindcă semnalele electromagnetice se propagă cu viteza luminii – de milioane de ori mai repede decât semnalele neuronale. însă cu cât o IA este mai mare, cu atât trebuie să fie mai lente gândurile ei globale, pentru a da timp informaţiei să circule între toate părţile sale, după cum am văzut în capitolul 4. Este aşadar de aşteptat ca o IA „Geea" de dimensiunea Pământului să aibă doar vreo zece experienţe conştiente pe secundă, la fel ca un om, iar o IA de dimensiunea unei galaxii ar avea doar un gând global cam la fiecare 100 000 de ani – adică nu mai mult de vreo 100 000 de gânduri în întreaga istorie a universului nostru! Asta ar face ca o IA mare să-şi delege calculele celor mai mici subsisteme capabile să le efectueze, la fel cum mintea noastră conştientă a delegat reflexul de a clipi unui subsistem mic, rapid şi neconştient. Am văzut că procesarea conştientă a informaţiei în creierul nostru pare să fie doar vârful unui aisberg neconştient, dar este de aşteptat să se ajungă şi mai departe pentru marile IA din viitor: dacă au o conştiinţă unică, atunci e probabil să ignore aproape întreaga procesare a informaţiei care are loc în ele. în plus, deşi experienţele conştiente de care se bucură pot fi extrem de complexe, acestea se desfăşoară în ritm de melc în comparaţie cu activităţile rapide ale părţilor mai mici.

Aceasta accentuează controversa legată de întrebarea dacă părţile unei entităţi conştiente pot fi şi ele conştiente.

IIT prezice că nu, ceea ce înseamnă că dacă o IA astronomic de mare din viitor e conştientă, atunci aproape întreaga procesare a informaţiei e neconştientă. Prin urmare, dacă o civilizaţie de IA mici îşi îmbunătăţeşte capacităţile de comunicare până în punctul în care apare o unică minte agregată conştientă, atunci conştiinţele ei individuale mult mai rapide dispar brusc. Pe de altă parte, dacă predicţia UT e falsă, mintea agregată poate coexista cu tot restul minţilor conştiente mai mici. N e putem chiar închipui o ierarhie de conştiinţe la toate nivelurile, de Ia microscopic Ia cosmic.

După cum am văzut mai sus, procesarea neconştientă a informaţiei în creierele noastre umane este legată de un mod de gândire rapid, automat, care nu necesită efort, numit de psihologi „Sistemul i".[3i] De pildă, Sistemul i ar putea da de ştire conştiinţei tale că extrem de complexa lui analiză asupra datelor input-ului vizual a dezvăluit că prietenul tău cel mai bun tocmai a sosit, fără să-ţi dea vreun indiciu despre felul în care a efectuat calculul. Dacă această legătură între sisteme şi conştiinţă se dovedeşte a fi valabilă, va fi tentant să generalizăm terminologia pentru IA, numind toate sarcinile rapide de rutină delegate subunităţilor neconştiente Sistemul i al IA. Gândirea globală lentă, controlată şi necesitând efort, presupunând că e conştientă, ar fi Sistemul 2 al IA. Noi, oamenii, avem experienţe conştiente implicând şi ceea ce voi numi „Sistemul 0": percepţia pasivă brută care are loc chiar şi atunci când stai fără să te mişti sau să te gândeşti, şi nu faci decât să observi lumea din jur. Complexitatea Sistemelor o, 1 şi 2 pare să crească progresiv, aşa încât e frapant faptul că doar cel din mijloc e neconştient. IIT explică asta spunând că informaţia senzo rială brută din Sistemul o e stocată în structuri cerebrale de tip reţea, cu un grad înalt de integrare, în vreme ce Sistemul 2 are un grad înalt de integrare datorită buclelor de feedback, în

care toate informaţiile de care eşti conştient acum pot influenţa stările viitoare ale creierului tău. Pe de altă parte, tocmai predicţia privind reţeaua conştientă a declanşat critica mai sus pomenită a lui Scott Aaronson la adresa IIT. Pe scurt, dacă o teorie care rezolvă problema destul de grea a conştiinţei va putea trece într-o bună zi o serie de teste experimentale riguroase, aşa încât să-i luăm în serios predicţiile, atunci se va îngusta mult spectrul opţiunilor pentru problema şi mai grea legată de ce ar putea simţi o IA conştientă din viitor.

Unele aspecte ale experienţei noastre subiective au în mod evident origini evolutive – de pildă, dorinţele noastre emoţionale legate de autoconservare (a mânca, a bea, a evita să fii ucis) şi de reproducere. Ar fi deci posibil să creăm o IA care să nu resimtă niciodată qualia precum foamea, setea sau dorinţa sexuală. După cum am văzut în capitolul precedent, dacă o IA extrem de inteligentă e programată să aibă practic orice obiectiv, oricât de ambiţios, s-ar strădui probabil să se autoconserve pentru a-l îndeplini. Dacă ar exista însă o societate de inteligenţe artificiale, acestora le-ar putea lipsi puternica noastră teamă de moarte: câtă vreme au copii ale lor însele, tot ce ar putea pierde ar fi amintirile acumulate de la ultimul backup, dacă au încredere că software-ul copiat va fi utilizat. în plus, capacitatea inteligenţelor artificiale de a copia cu uşurinţă informaţie şi software le-ar reduce probabil acel individualism caracteristic conştiinţei umane: diferenţa dintre tine şi mine ar scădea dacă ne-am putea partaja şi copia amintirile şi capacităţile, aşa încât un grup de inteligenţe artificiale învecinate s-ar simţi mai curând ca un unic organism cu o minte agregată.

Ar simţi oare o conştiinţă artificială că posedă liber-arbitru? Observaţi că deşi filozofii şi-au pus de mii de ani problema libe- rului-arbitru legat de *noi,* nu s-a ajuns la un consens nici măcar în privinţa felului în care trebuie pusă întrebarea.[33] Eu îmi pun o altă problemă, pesemne mai

uşor de abordat. Voi încerca să vă conving că răspunsul e pur şi simplu; „Da, orice entitate care ia decizii conştiente va *simţi* subiectiv că are liber-arbitru, indiferent dacă e biologică sau artificială.“ Deciziile se întind pe un spectru având aceste extreme:

1. Ştii exact de ce ai făcut o anumită alegere.
2. Habar n-ai de ce ai făcut o anumită alegere – simţi că ai ales la întâmplare sau după capriciu.

Discuţiile despre liberul-arbitru sunt de regulă concentrate în jurul încercării de a reconcilia comportamentul orientat spre obiective al celui care ia decizii cu legile fizicii. Dacă ai de ales între următoarele două explicaţii pentru ce ai făcut, care e cea corectă? „Am rugat-o să ne întâlnim, fiindcă îmi place foarte mult“ sau „Particulele care mă alcătuiesc au facut-o, mişcându-se conform legilor fizicii“. în capitolul precedent am văzut însă că *ambele* sunt corecte: ceea ce simţi că e comportament orientat spre obiective poate apărea din legile deterministe ale fizicii, lipsite de vreun obiectiv. Mai precis, când un sistem (creier sau IA) ia o decizie de tipul r, el calculează ce să hotărască folosind un algoritm determinist, iar motivul pentru care simte că el a hotărât este acela că într-adevăr el a luat hotărârea atunci când a calculat ce să facă. Mai mult, după cum a subliniat Seth Lloyd,[34] există în informatică o celebră teoremă care spune că pentru aproape toate calculele nu există un mod mai rapid de a determina rezultatul lor decât de a le face efectiv. Asta înseamnă că în general ţi-e imposibil să-ţi dai seama, într-un interval de timp mai mic de o secundă, ce vei hotărî peste o secundă să faci, ceea ce îţi întăreşte sentimentul că ai liber-arbitru. Dimpotrivă, când un sistem (creier sau IA) ia o decizie de tipul 2, el îşi programează pur şi simplu mintea să se bazeze în hotărâre pe rezultatul unui subsistem care acţionează ca un generator de numere aleatorii. în creiere şi în calculatoare, numerele aleatorii sunt cu uşurinţă

generate prin amplificarea zgomotului. Indiferent unde se află o decizie în spectrul de la 1 la 2, atât conştiinţele biologice, cât şi cele artificiale simt aşadar că au liber-arbitru: simt că într-adevăr ele sunt cele care hotărăsc şi nu pot prezice cu certitudine ce decizie vor lua până n-au încheiat procesul de gândire.

Unii oameni îmi spun că găsesc cauzalitatea degradantă, că privează de sens procesul gândirii şi îi reduce la statutul de „simple“ maşini Asemenea negativism mi se pare absurd şi neîntemeiat în primul rând, nu e nimic „simplu“ legat de creierul omului, care, după părerea mea, este cel mai complex obiect fizic din universul cunoscut. în al doilea rând, ce ar prefera ei? N-ar vrea ca propriile lor procese de gândire (calculele efectuate de creier) să fie acelea care iau hotărârile lor? Experienţa subiectivă a liberului-arbitru arată cum sunt resimţite calculele lor din interior; ei nu cunosc rezultatul calculului până nu-l încheie. Calculul *este* decizia.

SENSUL

Să ne întoarcem acum în punctul de pornire al acestei cărţi: cum vrem să fie viitorul vieţii? Am văzut în capitolul precedent că toate culturile îşi doresc un viitor plin de experienţe pozitive, dar nu ajungem la consens în privinţa a ceea ce e pozitiv şi a compromisului între ce e bine pentru diferite forme de viaţă. Asemenea controverse nu trebuie totuşi să ne abată atenţia de la elefantul din cameră: nu pot exista experienţe pozitive dacă nu există deloc experienţe – dacă nu există conştiinţă. Cu alte cuvinte, fără conştiinţă nu pot exista fericire, bine, frumos, sens sau rost, ci doar o risipă astronomică de spaţiu. Prin urmare, atunci când oamenii îşi pun întrebări despre sensul vieţii ca şi cum ar fi sarcina cosmosului să dea sens existenţei noastre, ei privesc lucrurile invers: *nu universul nostru dă sens fiinţelor conştiente, ci fiinţele conştiente dau sens universului nostru.* Principalul nostru obiectiv

pentru viitor ar trebui să fie menţinerea (şi, să sperăm, extinderea) conştiinţei biologice şi/sau artificiale în cosmosul nostru, în loc s-o ducem la pieire.

Dacă reuşim, ce vor simţi oamenii faţă de coexistenţa cu maşini tot mai deştepte? Vădita ascensiune inexorabilă a inteligenţei artificiale te nelinişteşte – şi dacă da, de ce? în capitolul 3 am văzut că ar fi relativ uşor ca o tehnologie folosind IA să ne satisfacă nevoi de bază cum ar fi securitatea şi bunăstarea materială, atâta vreme cât există voinţa politică pentru asta. Poate însă că a fi bine hrănit, îmbrăcat, adăpostit şi distrat ţi se pare insuficient. Dacă ni se garantează că IA se va ocupa de toate dorinţele şi nevoile noastre practice, nu vom sfârşi oare prin a simţi că viaţa noastră e lipsită de sens şi de scop, că vom trăi la fel ca animalele bine îngrijite de la grădina zoologică?

Tradiţional, noi, oamenii, ne-am întemeiat orgoliul pe ideea de *excepţionalism uman:* convingerea că suntem entităţile cele mai deştepte de pe planetă, aşadar unici şi superiori. Ascensiunea IA ne va obliga să devenim mai modeşti. Ideea de superioritate faţă de alţii (indivizi, grupuri etnice, specii etc.) a provocat mari neno rociri, şi a sosit pesemne momentul să renunţăm la ea. Nu numai

că excepţionalismul uman e responsabil de suferinţe din trecut, dar el nici nu-i necesar pentru înflorirea umanităţii: dacă am descoperi o civilizaţie extraterestră paşnică mult mai avansată decât noi în ştiinţă, artă şi toate celelalte lucruri care ne interesează, asta nu i-ar împiedica pe oameni să simtă că viaţa lor are un sens şi un scop. Am rămâne cu familiile, prietenii şi comunităţile mai largi, cu toate activităţile care ne dau un sens şi un scop, pierzând, să sperăm, doar aroganţa.

Din moment ce vorbim despre viitor, să ne gândim nu numai la sensul vieţilor noastre, dar şi la sensul universului nostru. în această privinţă, doi dintre fizicienii pe care îi admir cel mai mult, Steven Weinberg şi Freeman Dyson, au vederi diametral opuse. Weinberg, care a primit Premiul

Nobel pentru contribuţii fundamentale la modelul standard din fizica particulelor, a spus: „Cu cât universul pare mai comprehensibil, cu atât pare mai lipsit de sens.“[35] Pe de altă parte, după cum am văzut în capitolul 6, Dyson e mult mai optimist: deşi este de acord că universul *a fost* lipsit de sens, el crede că viaţa îi oferă acum tot mai mult sens, iar viitorul va arăta şi mai bine dacă viaţa va reuşi să se răspândească prin cosmos. Dyson îşi încheie astfel celebrul articol din 1979: „Care univers e mai aproape de adevăr, al Iui Weinberg sau al meu? într-obună zi, nu peste multă vreme, vom afla.“[36] Dacă universul nostru va cădea într o stare de eternă neconştienţă pentru că vom distruge viaţa de pe Pământ sau vom permite unei IA zombi să preia controlul asupra universului nostru, atunci Weinberg va avea câştig de cauză.

Privind din această perspectivă, ne dăm seama că, deşi ne-am concentrat în cartea de faţă asupra viitorului inteligenţei, viitorul conştiinţei e şi mai important, fiindcă el oferă un sens. Filozofii apelează la latină pentru a pune în antiteză *sapientia* (capacitatea de a gândi inteligent) şi *sententia* (capacitatea de a percepe subiectiv qualia). Noi, oamenii, ne-am construit identitatea pornind de la faptul că suntem *Homo sapiens,* cele mai deştepte entităţi de prin părţile noastre. Cum suntem pe cale de a fi întrecuţi de maşini şi mai deştepte, propun să ne redefinim ca *Homo sentiend*

REZUMAT:

• Nu există o definiţie indiscutabilă a „conştiinţei“. Eu folosesc termenul într-un sens larg, neantropocentric: *conştiinţă = experienţă subiectivă.*

• Dacă inteligenţele artificiale sunt conştiente în acest sens este ceea ce contează pentru cele mai spinoase probleme etice şi filozofice puse de ascensiunea IA: Pot suferi inteligenţele artificiale? Ar trebui ele să aibă drepturi? Este uploadarea o sinucidere subiectivă? Ar putea fi un cosmos din viitor plin de lAapocalipsa zombi absolută?

• Problema înţelegerii inteligenţei n-ar trebui să fie amestecată cu trei probleme separate legate de conştiinţă: „problema destul de grea“ de a prezice care sisteme fizice sunt conştiente, „problema şi mai grea“ de a prezice qualia şi „problema *cu adevărat* grea“ – de ce ajunge ceva

să fie conştient.
• „Problema destul de grea“ a conştiinţei face obiectul ştiinţei, din moment ce o teorie care prezice care dintre procesele tale cerebrale sunt conştiente e testa bilă experimental şi falsificabilă, însă deocamdată nu e limpede cum ar putea ştiinţa rezolva pe deplin cele două probleme mai grele.
• Experimentele neuroştîinţifice sugerează că multe comportamente şi regiuni ale creierului sunt neconştiente, o mare parte din experienţa noastră conştientă reprezentând un rezumat post factum al unei mult mai vaste cantităţi de informaţie neconştientă.
• Generalizarea predicţiilor privind conştiinţa de la creiere la maşini necesită o teorie. Conştiinţa pare să aibă nevoie nu de un tip anume de particule sau de câmpuri, ci de un tip anume de procesare a informaţiei, în mare măsură autonom şi integrat, aşa încât sistemul în ansamblu să fie relativ autonom, dar nu şi părţile lui.
• Conştiinţa se simte pesemne atât de non-fizică pentru că este de două ori independentă de substrat: dacă ea este felul în care simte informaţia atunci când este procesată în anumite moduri complexe, ceea ce contează este doar structura procesării informaţiei, nu şi structura materiei care efectuează procesarea informaţiei.
• Dacă e posibilă conştiinţa artificială, atunci spaţiul tuturor experienţelor posibile ale IA este probabil uriaş în comparaţie cu cel al experienţelor umane, întins pe un spectru vast de qualia şi scări temporale – toate împărtăşind sentimentul că posedă liber-arbitru.
• Din moment ce ne putem pierde orice sens în absenţa conştiinţei, nu universul nostru dă sens fiinţelor conştiente, ci fiinţele conştiente dau sens universului nostru.
• Aceasta sugerează că, pe măsură ce ne pregătim să fim întrecuţi de maşini tot mai deştepte, ne consolează mai curând faptul că suntem *Homosentiens,* decât că suntem *Homo sapiens.*

EPILOG

Povestea Echipei FLI

Cel mai trist lucru legat de viaţa din prezent este ci ştiinţa acumulează cunoaştere mai rapid decât acumulează societatea înţelepciune. – Isaac Asimov

lată-ne, dragă cititorule, ajunşi Ia sfârşitul cărţii, după ce am explorat originea şi destinul inteligenţei, obiectivelor şi sensului. Cum am putea aşadar transforma aceste idei în acţiune? Ce ar trebui să facem concret pentru a ne fauri un viitor cât mai bun? Este întrebarea pe care mi-o pun chiar acum, pe 9 ianuarie 2017, când mă aflu pe scaunul de lângă fereastră, în drumul de întoarcere de la San Francisco la Boston, după conferinţa despre IA pe care am organizat-o Ia Asilomar, aşa că permiteţi-mi să închei această carte împărtăşindu-vă gândurile mele.

Meia doarme alături de mine, după ce a petrecut multe nopţi cu pregătirile şi organizarea. Ce săptămână agitată am avut! Am reuşit să-i aducem laolaltă pe aproape toţi cei menţionaţi în cartea de faţă, pentru câteva zile, la această continuare a conferinţei din Puerto Rico – între ei, oameni de afaceri ca Elon Musk şi Larry Page, cercetători IA de vârf din universităţi şi companii cum sunt DeepMind, Google, Facebook, Apple, IBM, Microsoft şi Baidu, precum şi economişti, jurişti, filozofi şi alţi gânditori remarcabili (vezi figura 9.1). Rezultatele au depăşit până şi speranţele mele cele mai cutezătoare, şi sunt atât de optimist în privinţa viitorului vieţii cum n-am mai fost de mult. în acest epilog am să explic de ce.

S-A NĂSCUT FLI

De la vârsta de 14 ani, când am aflat de cursa înarmărilor nucleare, mă îngrijorează ideea că puterea tehnologiei noastre creşte mai rapid decât înţelepciunea cu care o controlăm. M-am hotărât aşadar să introduc un capitol despre această provocare în prima mea carte, *Our Mathematical Universe,* deşi restul ei se ocupa în principal de fizică. în 2014, de Anul Nou, am luat hotărârea că nu mai am voie să deplâng această situaţie fără să reflectez serios la ce pot face eu însumi în privinţa asta, şi am respectat legământul în cursul turneului de promovare a cărţii mele din ianuarie: Meia şi cu mine ne am gândit mult

cum putem înfiinţa un soi de organizaţie nonprofit concentrată pe ameliorarea viitorului vieţii prin intermediul tehnologiei.

Ea a insistat să-i dăm un nume pozitiv, în nici un caz ceva de genul „Institutul Apocalipsei şi întunericului*' sau „Institutul Hai-să-ne-temem-de-viitor“. Din moment ce Institutul pentru Viitorul Omenirii (Future of Humanity Institute) exista deja, am ales numele de Institutul pentru Viitorul Vieţii (Future of Life Institute – FLI), care avea avantajul de a fi mai cuprinzător. Pe 22 ianuarie turneul ne-a adus la Santa Cruz, iar pe când soarele californian apunea deasupra Pacificului, am luat cina cu vechiul nostru prieten Anthony Aguirre şi l-am convins să ni se alăture. Nu e numai unul dintre cei mai inteligenţi şi mai idealişti oameni pe care îi cunosc, dar el a reuşit să conducă, împreună cu mine, timp de un deceniu şi o altă organizaţie nonprofit, Institutul pentru întrebările Fundamentale (Foundational Questions Institute – vezi http://fqxi.org).

în săptămâna următoare, turneul de promovare a cărţii m-a adus la Londra. Fiindcă mă preocupa viitorul IA, l-am căutat pe Demis Hassabis, care a avut amabilitatea să mă invite să vizitez sediul central al DeepMind. Am fost uimit să descopăr cât se dezvoltase de când Demis mă vizitase Ia MIT, cu doi ani în urmă, când Google tocmai cumpărase DeepMind cu 650 de milioane de dolari. Văzând acum birourile încăpătoare pline de minţi strălucite care urmăreau obiectivul îndrăzneţ al lui Demis de a „rezolva problema inteligenţei“, am avut sentimentul că succesul era într-ade- văr posibil.

A doua zi de dimineaţă am vorbit cu prietenul meu ţaan Tallinn prin Skype, software-ul la crearea căruia contribuise. I-am explicat cum vedem noi proiectul FU, iar peste o oră a hotărât să ne finanţeze cu 100 ooo de dolari pe an! M-a impresionat că cineva poate avea atâta

încredere în mine, mai ales că un an mai târziu, după conferinţa din Puerto Rico menţionată în capitolul i, mi-a spus, râzând, că a fost cea mai bună investiţie din câte a făcut vreodată.

A doua zi, editorul meu mi-a lăsat o pauză în program, de care am profitat vizitând Muzeul Ştiinţei din Londra. Fiind de atâta timp obsedat de trecutul şi viitorul inteligenţei, am simţit brusc că mă plimbam prin concretizarea fizică a gândurilor mele. Este adunată acolo o colecţie fantastică de exponate reprezentând sporirea cunoaşterii umane, de la locomotiva Rocket a lui Stephenson la Modelul T al lui Ford, o replică în mărime naturală a modulului lunar al lui Apollo ir şi computere de la calculatorul mecanic al lui Babbage „Mecanismul Diferenţial“ până la cele din zilele noastre. Este prezentată şi evoluţia înţelegerii minţii, de la experimentele lui Galvani cu picioare de broască până la neuroni, EEG şi IRMf.

îmi vine foarte rar să plâng, dar asta am făcut când am ieşit din muzeu – într-un tunel aglomerat, în drum spre staţia de metrou South Kensington. Cei din jur îşi vedeau fiecare de drumul lui, din fericire fără să bănuiască la ce mă gândeam. La început, noi, oamenii, am descoperit cum să reproducem anumite procese naturale cu maşini, generând vânt, lumină şi cai-putere pe cale mecanică. Treptat, am înţeles că şi trupurile noastre sunt maşini. Apoi, descoperirea celulei nervoase a început să estompeze graniţa dintre trup şi minte. Am construit maşini care întrec nu numai muşchii, dar şi minţile noastre. Şi atunci, în paralel cu descoperirea a ceea ce suntem, nu cumva în mod inevitabil devenim desueţi? Ar fi tragic.

Acest gând m-a înspăimântat, dar mi-a şi întărit hotărârea de a mă ţine de legământul făcut de Anul Nou. Simţeam că, pentru

Figura 9.1 *Conferința* noastră Asilomar din ianuarie 2017, continuarea celei din Puerto Rico, a adus laolată un remarcabil grup de cercetători din IA şt domeniile înrudite. Rândul din spate, de la stânga la dreapta: Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin, Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán *Ó* hÉigeartaigh, Owaín Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Rrth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey, ?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferriera, János Kramar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Roth berg. Rândul din faţă: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan Hadfield-Menell, Verity Harding, Shivon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas,

Roman Yampolskiy, Marin So lj ai ic, Lawrence Krauss, Stuart Russell, Eric Bryn- jolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, MaxTegmark, Adrian Weller, Jose Hernan- dez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Go rd on-Levitt. Au sosit mai târziu: Guru Banavar, Demis Hassabis, Rao Kam- bhampati, Elon Musk, Larry Page, Anthony Romero

a completa echipa de fondatori FLI, mai aveam nevoie de cineva care să conducă un grup de tineri voluntari idealişti. Alegerea firească a fost Viktoriya Krakovna, o strălucită absolventă de Harvard care, pe lângă faptul că fusese medaliată cu argint la Olimpiada Internaţională de Matematică, înfiinţase Citadela, unde se aflau o duzină de tineri idealişti care voiau sase afirme. Cinci zile mai târziu, Meia şi cu mine am invitat-o la noi pentru a-i vorbi despre planul nostru, iar înainte să terminăm de mâncat sushi-ul FLI se născuse.

AVENTURA DIN PUERTO RICO

Aşa a început o aventură uimitoare, care continuă şi azi. După cum am menţionat în capitolul i, găzduim cu regularitate la noi acasă întâlniri de *brainstorming,* la care participă zeci de studenţi, profesori şi persoane interesate

de subiect din zona Bostonului,

Figura 9.2 Jaan Tallinn, Anthony Aguirre, al dumneavoastră MaxTegmark, Mela Chita-Tegmaric şi Viktoriya Krakovna sărbătorind cu sushi înfinţarea FU pe 23 mai 2014.

ideile cele mai apreciate fiind transformate în proiecte – primul a fost acel editorial despre care am vorbit în capitolul 1, avându-i drept coautori pe Stephen Hawking, Stuart Russell şi Frank Wilczek, care a iniţiat dezbaterea publică. în paralel cu primii paşi pentru înfiinţarea unei noi organizaţii (angajări, alegerea unui comitet consultativ şi lansarea unui site), a avut loc un eveniment inaugural în faţa unui numeros public de la MIT, la care Alan Aida a explorat viitorul tehnologiei împreună cu experţi de primă mărime.

Ne-am concentrat în restul anului asupra conferinţei din Puerto Rico, care, aşa cum am menţionat în capitolul 1, urmărea să-i atragă pe cei mai importanţi cercetători IA din lume într-o dezbatere privind felul în care LA ar putea deveni benefică. Obiectivul nostru era de a deplasa discuţia despre siguranţa IA de la temeri Ia acţiune, de la dispute generate de îngrijorare la acordul asupra unor proiecte concrete de cercetare care ar putea maximiza imediat şansele unui rezultat pozitiv. în pregătirea conferinţei, am adunat de pretutindeni idei promiţătoare despre

siguranţa IA, primind reacţia comunităţii la lista noastră tot mai bogată. Cu ajutorul lui Stuart Russell şi al unui grup de tineri voluntari devotaţi, între care Daniel Dewey, János Krámar şi Richard Mallah, am distilat aceste priorităţi de cercetare într-un docu- ment care să fie discutat la conferinţă.[1] Obţinerea unui consens privind faptul că era mult de lucru în domeniul siguranţei IA speram să-i încurajeze pe cercetători să înceapă această muncă. Triumful ultim ar fi fost să convingem pe cineva să finanţeze asemenea cercetări, fiindcă până atunci ele nu primiseră practic nici un sprijin din partea agenţiilor de finanţare guvernamentale.

Şi iată că îşi face apariţia Elon Musk. Pe 2 august, ne a atras atenţia celebrul său mesaj lansat pe tweeter: „Merită citită *Superintelligence de* Bostrom. Trebuie să fim superprecauţi cu IA. Potenţial mai periculoasă decât armele nucleare.“ A aflat despre eforturile noastre, şi am vorbit cu el la telefon câteva săptămâni mai târziu. Deşi mă simţeam foarte emoţionat şi intimidat, rezultatul a fost remarcabil: a acceptat să se alăture comitetului consultativ FLI, să participe la conferinţă şi eventual să finanţeze primul program de cercetare în domeniul siguranţei IA, care avea să fie anunţat în Puerto Rico. Asta ne-a entuziasmat şi ne-a făcut să ne sporim eforturile pentru a organiza o conferinţă excelentă, a identifica teme de cercetare promiţătoare şi a obţine pentru acestea sprijinul comunităţii.

Peste două luni am ajuns în fine să-l întâlnesc pe Elon, sosit la MIT pentru un simpozion. M-am simţit ciudat să mă aflu singur cu el într-o încăpere mică, la doar câteva momente după ce fermecase o mie de studenţi de la MIT ca un star rock, însă apoi nu m am putut gândi decât la proiectul nostru comun. L-am plăcut din prima clipă. Radia sinceritate şi m-a impresionat cât de preocupat era de viitorul pe termen lung al omenirii şi cu câtă temeritate îşi transforma aspiraţia în acţiune. Dorea ca omenirea să

exploreze şi să colonizeze universul nostru, aşa că a înfiinţat o companie spaţială. Dorea energie fără poluare, aşa că a înfiinţat o companie care să producă energie solară şi una care să fabrice maşini electrice. înalt, chipeş, bun vorbitor şi cunoscând o mulţime de lucruri, nu era de mirare că lumea îl asculta cu sufletul la gură.

Din păcate, acest eveniment de la MIT m-a făcut să-mi dau seama ce rol nefast pot juca uneori mediile de informare. Elon a ţinut timp de o oră un discurs fascinant despre explorarea spaţiului, care putea fi un formidabil spectacol TV. La urmă, un student i-a pus o întrebare în afara subiectului despre IA. în răspunsul lui, Elon a spus, între altele, „cu inteligenţa artificială invocăm demonul" – şi^uru/lucru pe care majoritatea mediilor de informare l-au preluat, şi, în general, scos din context. M-a frapat faptul că mulţi jurnalişti făceau în mod inconştient *exact opusul* a ceea ce încercam noi să facem în Puerto Rico. Dacă noi voiam să obţinem un consens în rândul comunităţii, subliniind ce ne uneşte, mediile de informare aveau interesul să sublinieze divizarea. Cu cât puteau anunţa o controversă mai aprigă, cu atât mai mare erau ratigul şi veniturile lor. în plus, în vreme ce noi voiam să reunim oameni din întreg spectrul opiniilor, pentru ca măcar să se înţeleagă unul pe altul mai bine, relatările din mediile de informare îi întărâtau pe cei cu opinii diferite pe unii împotriva altora şi alimentau neînţelegerile publicând doar citatele cele mai provocatoare scoase din context. De aceea am hotărât să nu permitem accesul jurnaliştilor la întâlnirea din Puerto Rico şi am impus „Regula de la Chatham House"††††††††††††, care interzice participanţilor să

†††††††††††††† Regulă stabilită în 1927 la Institutul Regal pentru Afaceri Interne (al cărui sediu e Chatham House) care permite participanţilor la o întrunire să folosească informaţiile căpătate cu acest prilej, dar n-au voie să dezvăluie identitatea şi afilierea persoanelor care au oferit acele informaţii, şi nici identitatea celorlalţi participanţi. (N. t.)

dezvăluie ulterior cine ce a spus.‡‡‡‡‡‡‡‡‡‡‡‡

Până la urmă, conferinţa noastră din Puerto Rico s-a dovedit a fi un succes, dar n-a fost simplu. Apropierea termenului-limită ne-a obligat să ne înteţim eforturile. De pildă, *a* trebuit să vorbesc la telefon sau prin skype cu mulţi cercetători IA pentru a aduna o masă critică de participanţi care să-i atragă şi pe alţii, iar momentele dramatice n-au lipsit – ca atunci când m-am trezit la ora şapte dimineaţa, pe 27 decembrie, pentru a-l suna pe Elon, aflat în Uruguay, şi a-l auzi, printr-o legătură telefonică proastă, că-mi spune: „Nu cred că o să meargă.“ Era îngrijorat că un program de cercetare în domeniul siguranţei IA ar putea genera un fals sentiment de securitate, permiţând cercetătorilor recalcitranţi să-şi continue munca neabătut, pretinzând de formă că-i preocupă siguranţa. Apoi însă, deşi legătura se întrerupea mereu, am vorbit pe larg despre imensele avantaje care s-ar obţine dacă am aduce subiectul în centrul atenţiei şi dacă tot mai mulţi cercetători ar lucra în domeniul siguranţei IA. După ce convorbirea s-a întrerupt definitiv, mi-a trimis unul dintre cele mai frumoase mesaje pe care le-am primit vreodată: „Am pierdut legătura. Oricum, documentele arată bine. Mă bucur să susţin cercetarea cu $şM timp de trei ani. Sau poate $ioM?“

Patru zile mai târziu, anul 2015 a început bine pentru Meia şi pentru mine – înainte de conferinţă, am avut răgazul să ne relaxăm şi am dansat pe o plajă din Puerto Rico luminată de artificii. Şi conferinţa a avut un început foarte bun: s-a ajuns la un remarcabil consens privind faptul că e nevoie de mai multă cercetare în domeniul

‡‡‡‡‡‡‡‡‡‡‡‡‡ Această experienţă m-a făcut să mă gândesc mai atent la felul în care ar trebui eu însumi să interpretez ştirile. Deşi, evident, fusesem conştient că majoritatea informaţiilor sunt orientate politic, mi-am dat seama acum că ele pot fi tendenţioase chiar şi atunci când e vorba de alte subiecte decât politica. *(N. a.)*

siguranţei IA, iar prin contribuţiile participanţilor, acel document despre priorităţile în cercetare la care lucraserăm atât de intens a fost îmbunătăţit şi a căpătat forma finală. Am făcut să circule scrisoarea deschisă în favoarea cercetărilor pentru siguranţa IA, şi am fost încântat că l-au semnat aproape toţi.

Meia şi cu mine am avut o minunată întâlnire în camera noastră de hotel cu Elon, care şi-a dat acordul asupra planurilor detaliate pentru programul nostru de burse. Pe Meia au impresionat-o felul simplu şi deschis cu care vorbea despre viaţa lui personală şi interesul pe care ni-l arăta. A întrebat cum ne-am cunoscut şi a ascultat cu plăcere povestea ei amănunţită. A doua zi am filmat un interviu cu el despre siguranţa IA şi despre motivele pentru care voia s-o sprijine în tot ce părea să aibă şanse de reuşită?

Punctul culminant al conferinţei, anunţul donaţiei lui Elon, era programat pentru duminică 4 ianuarie 2015, la ora şapte seara. iar eu eram atât de agitat, încât m-am foit în pat toată noaptea dinaintea acelei zile. Apoi, cu un sfert de oră înainte să plecăm spre locul unde trebuia să se desfăşoare sesiunea, am primit o lovitură! Asistentul lui Elon ne a sunat şi ne-a spus că s-ar părea că Elon nu poate face anunţul. Meia mi-a mărturisit că nu mă văzuse niciodată atât de nervos şi dezamăgit. Elon a apărut în cele din urmă, iar eu, în timp ce stăteam de vorbă cu el, număram secundele până la începerea sesiunii. A explicat că mai erau doar două zile până la lansarea rachetei SpaceX, sperând ca pentru prima dată o navă-dronă să aterizeze cu succes, era deci un moment crucial, drept urmare echipa SpaceX nu voia ca mediile de informare să distragă atenţia publicului cu alte ştiri implicându-l pe el. Anthony Agurre, calm şi echilibrat ca de obicei, a subliniat că *nimeni* nu-şi dorea distragerea atenţiei, nici Elon, nici comunitatea IA. Am ajuns cu câteva minute întârziere la

sesiunea al cărei moderator eram, dar aveam un plan: nu va fi pomenită nici o sumă, ca să ne asigurăm că anunţul nu e incitant pentru presă, şi le voi cere tuturor să respecte regula de la Chatham House păstrând secretul asupra anunţului Iui Elon timp de nouă zile dacă racheta lui ajunge la staţia spaţială, indiferent de reuşita sau nereuşita aterizării; el a spus că va fi nevoie de mai mult timp dacă racheta explodează la lansare.

Numărătoarea inversă pentru anunţ a ajuns în fine la zero. Partenerii mei de discuţie se aflau în continuare pe scenă, alături de mine: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg şi Vemon Vinge. Treptat, aplauzele publicului s-au stins, dar colegii mei erau mai departe pe scenă, fiindcă le spusesem să rămână, fără să le explic de ce. Meia avea să-mi spună că pulsul ei o luase razna şi strângea pe sub masă mâna Viktoriyei Krakovna ca să se calmeze. Zâmbeam, ştiind că sosise momentul mult aşteptat şi mult sperat.

Am spus că eram bucuros că se ajunsese la un consens privind necesitatea de a spori cercetările pentru ca IA să fie benefică, şi că erau multe direcţii de cercetare pe care ar trebui să le urmăm de îndată. S-a vorbit însă în această sesiune şi despre riscuri serioase, am adăugat, aşa încât ar fi bine să ne ridicăm moralul înainte de a merge Ia banchetul conferinţei. „Aşadar îi dau microfonul lui... Elon Musk!“ Am simţit că trăiesc un moment istoric când Elon a luat microfonul şi a anunţat că va dona o sumă importantă pentru cercetările în siguranţa IA. Nu-i de mirare că au izbucnit aplauze furtunoase. După cum plănuiserăm, n-a menţionat suma, dar eu ştiam că erau zece milioane de dolari, aşa cum ne înţeleseserăm.

După conferinţă, Meia şi cu mine am fost în vizită la părinţii noştri în Suedia şi România, şi am urmărit cu emoţie la Stockholm, împreună cu tatăl meu, lansarea

rachetei. Din păcate, încercarea de aterizare s-a încheiat cu ceea ce Elon numeşte eufemistic RUD, „rapid unscheduled disassembly“ (dezasamblare rapidă neprogramată) – echipa sa avea să obţină o amerizare reuşită cincisprezece luni mai târziul Toţi sateliţii au fost însă lansaţi cu succes pe orbită, la fel ca programul nostru de burse anunţat de Elon pe tweeter milioanelor sale de admiratori.[4]

SIGURANŢA IA AJUNGE ÎN CENTRUL ATENŢIEI

Un obiectiv esenţial al conferinţei din Puerto Rico fusese acela de a aduce în centrul atenţiei cercetările din domeniul siguranţei IA, şi am urmărit cu însufleţire această evoluţie în mai multe etape. Mai întâi a fost însăşi întrunirea, la care numeroşi cercetători s-au bucurat să vadă că, implicându-se în acest domeniu, fac parte dintr-o comunitate în continuă creştere. M-au impresionat încurajările primite de la ei. De pildă, profesorul Bart Selman de Ia Universitatea Corneli mi-a trimis un e-mail în care spunea: „Vă mărturisesc cu toată sinceritatea că n-am văzut niciodată o întrunire ştiinţifică atât de bine organizată, atât de pasionantă şi de stimulativă intelectual.“

Următoarea etapă a început pe n ianuarie, când Elon a lansat pe tweeter mesajul „Dezvoltatori de vârf din inteligenţa artificială semnează o scrisoare deschisă prin care susţin necesitatea cercetărilor în siguranţa IA“,[5] ataşând link ul unei pagini cu semnături, unde au apărut curând peste opt mii de nume, lista conţinând multe dintre figurile proeminente din IA. Brusc, a devenit mai greu să susţii că cei preocupaţi de siguranţa IA nu ştiau despre ce vorbesc, fiindcă asta înseamnă acum să spui că marii specialişti în IA nu ştiau despre ce vorbesc. Scrisoarea deschisă a fost preluată de mediile de informare din

întreaga lume, şi ne am dat atunci seama că procedaserăm bine interzicându-le jurnaliştilor accesul la conferinţă. Deşi cuvântul cel mai alarmant din scrisoare era „capcane“, acesta a declanşat în presă titluri precum „Elon Musk şi Stephen Hawking semnează o scrisoare deschisă în speranţa de a împiedica revolta roboţilor“, ilustrate de terminatori criminali. Dintre sutele de articole pe care le-am găsit, preferatul nostru a fost unul care îşi bătea joc de celelalte, spunând că „un titlu care evocă androizi călcând în picioare cranii umane face din tehnologia complexă care poate transforma lumea un spectacol de bâlci“.[6] Din fericire, au fost şi multe articole obiective, iar acum trebuia să facem faţă torentului de semnături, să le verificăm manual pentru a ne proteja credibilitatea şi a elimina nume apărute în batjocură, cum ar fi „HAL 9000“, „Terminator“, „Sarah Janette Connor“ sau „Skynet“ Pentru aceasta şi pentru viitoarele scrisori deschise, Viktoriya Krakovna şi János Krámar au organizat o brigadă de verificatori voluntari, care-i includea pe Jesse Galei, Eric Gastfriend şi Ravathi Vinoth Kumar, lucrând în schimburi, astfel încât atunci când Ravathi se ducea la culcare în India, Eric prelua ştafeta la Boston, şi aşa mai departe.

A treia etapă a început patru zile mai târziu, când Elon a postat pe tweeter un link la anunţul nostru că dona zece milioane de dolari pentru cercetări în siguranţa LA.[7] Peste o săptămână am lansat un portal online unde cercetători din întreaga lume puteau face cereri de finanţare. Am putut crea sistemul de cereri atât de repede fiindcă Anthony şi cu mine ne ocupaserăm de zece ani cu asemenea competiţii pentru burse în fizică. Open Philantropy Project, o instituţie filantropică din California orientată spre finanţări de mare impact, a avut generozitatea să adauge o sumă la cea donată de Elon, aşa încât puteam oferi mai multe burse. Nu bănuiam câte cereri vor fi, din moment ce domeniul era nou, iar până la termenul-limită mai rămânea puţin timp. Răspunsul a fost

copleşitor – vreo trei sute de echipe din lumea întreagă cereau aproximativ o sută de milioane de dolari. O comisie de profesori şi cercetători IA a analizat propunerile şi a ales 37 de echipe câştigătoare, care au fost finanţate pe o perioadă de până la trei ani. Când am anunţat lista câştigătorilor, presa a răspuns pentru prima dată nuanţat, fără imagini cu roboţi ucigaşi. Devenea în fine evident că siguranţa IA nu era vorbărie goală, ci o activitate utilă, iar multe mari echipe de cercetare s-au alăturat acestui efort.

A patra etapă s-a desfăşurat organic în cursul următorilor doi ani, cu numeroase publicaţii tehnice şi zeci de seminarii despre siguranţa IA în întreaga lume, ţinute de regulă pe lângă cele mai importante conferinţe dedicate inteligenţei artificiale. Specialişti perseverenţi încercaseră de mulţi ani să angajeze comunitatea IA în cercetări legate de siguranţă, cu succes limitat, dar acum mişcarea luase cu adevărat avânt. O mare parte din acele publicaţii au fost finanţate prin programul nostru de burse, iar FU a depus eforturi pentru a organiza şi finanţa cât mai multe dintre acele seminarii, însă numărul lor a crescut şi graţie cercetătorilor IA care şi-au dedicat timpul şi resursele acestui scop. Ca atare, şi mai mulţi specialişti au aflat despre cercetările legate de siguranţă de la colegii lor, descoperind că, în afară de faptul că e utilă, această activitate poate fi şi foarte atractivă, implicând probleme matematice şi computaţionale interesante.

Desigur, nu toţi găsesc atractive ecuaţiile complicate. La doi ani după conferinţa din Puerto Rico, am prefaţat conferinţa noastră Asilomar cu un seminar tehnic la care câştigătorii burselor FLI şi-au prezentat rezultatele, pe un ecran mare fiind proiectate diapozitive pline cu simboluri matematice. Moshe Vardi, profesor de IA la Universitatea Rice, a spus în glumă că şi-a dat seama că am reuşit să impunem ca domeniu de cercetare siguranţa IA atunci când întâlnirile au devenit plictisitoare.

Creşterea spectaculoasă a activităţii legate de siguranţa IA nu s-a limitat la mediul academic. Amazon, DeepMind, Facebook, Google, IBM şi Microsoft au lansat un parteneriat industrial pentru IA benefică.[8] Noi donaţii importante pentru siguranţa IA au permis extinderea cercetărilor la cele mai mari organizaţii nonprofit înrudite cu a noastră: Institutul de Cercetare pentru Inteligenţa Maşinilor de la Berkeley, Institutul pentru Viitorul Omenirii de la Oxford şi Centrul de Studii privind Riscurile la care Este Expusă Existenţa de la Cambridge. Donaţii ulterioare de peste zece milioane de dolari au iniţiat şi alte eforturi îndreptate spre IA benefică: Centrul Leverhulme pentru Viitorul Inteligenţei de la Cambridge, Fundaţia Gates K&L pentru Etică şi Tehnologii Computaţionale din Pittsburg şi Fondul pentru Etica şi Administrarea Inteligenţei Artificiale din Miami. în fine, dar nu în ultimul rând, angajându-se să participe cu un miliard de dolari, Elon Musk a intrat în parteneriat cu alţi oameni de afaceri pentru a lansa OpenAI, o companie nonprofit din San Francisco având drept obiect inteligenţa artificială benefică. Cercetarea legată de siguranţa IA se impusese definitiv.

în paralel cu cercetarea, au sporit şi opiniile exprimate, atât individual, cât şi colectiv. Parteneriatul industrial pentru IA şi-a publicat principiile fondatoare, iar lungi rapoarte cu liste de recomandări au fost publicate de guvernul Statelor Unite, Universitatea Stanford şi IEEE (cea mai mare organizaţie profesională din lume), împreună cu alte zeci de rapoarte şi luări de poziţie din diverse surse.[9]

Am urmărit să stimulăm discuţiile semnificative în rândul participanţilor la Asilomar şi să aflăm asupra căror puncte, dacă există vreunele, cade de acord această comunitate eterogenă. Lucas Perry şi-a asumat aşadar sarcina eroică de a citi toate acele documente pe care le-

am găsit şi de a extrage opiniile prezentate în ele. într-un efort-maraton iniţiat de Anthony Aguirre şi încheiat printr-o serie de lungi teleconferinţe, echipa noastră FLI a încercat să grupeze laolaltă opiniile similare şi să elimine verbiajul birocratic redundant, pentru a obţine o listă de principii succinte, incluzând şi opiniile nepublicate, dar semnificative, exprimate în discuţii informale şi cu alte prilejuri. Dar lista conţinea încă ambiguităţi

Figura 9.3 Grupuri de specialişti reflectând asupra principiilor IA la Asilomar.

sau contradicţii, şi lăsa loc interpretărilor, aşa încât cu o lună înainte de conferinţă am trimis-o participanţilor, primind opiniile şi sugestiile lor legate de principii noi sau îmbunătăţite. Din răspunsul comunităţii a rezultat o listă a principiilor semnificativ revizuită, pentru a fi folosită la conferinţă.

La Asilomar, lista a fost mai departe îmbunătăţită în două etape. La început, mici grupuri au discutat principiile care le interesau cel mai mult (figura 9.3), au făcut ajustări de detaliu şi au adăugat noi principii şi versiuni alternative ale unora vechi. în cele din urmă, am efectuat un sondaj în rândul participanţilor pentru a determina susţinerea de care se bucură fiecare versiune a fiecărui principiu.

Acest proces colectiv a fost deopotrivă exhaustiv şi epuizant, Anthony, Meia şi cu mine luându-ne din timpul de somn şi de masă ca să compilăm tot ce era necesar pentru următoarele etape. A fost însă şi palpitant. După asemenea discuţii amănunţite, complicate, uneori aprige, obţinând o gamă largă de reacţii, am fost uimiţi de nivelul ridicat al consensului la ultimul sondaj în privinţa multor principii, unele obţinând o susţinere de peste 97%. Acest consens ne-a permis să aşezăm foarte sus ştacheta

exigenţei pentru lista finală: am păstrat doar principiile cu care au fost de acord cel puţin 90% dintre participanţi. Deşi asta a însemnat ca unele principii întrunind o largă apreciere să fie abandonate în ultimul moment, între care şi câteva dintre preferatele mele,[10] a permis ca majoritatea participanţilor să le aprobe pe toate cu inima împăcată, semnând un document care a circulat printre ei. Iată în continuare rezultatul.

PRINCIPIILE IA DE LA CONFERINŢA ASILOMAR

Inteligenţa artificială a pus deja la dispoziţie instrumente benefice care sunt zilnic folosite de oameni din lumea întreagă. Dezvoltarea ei continuă, călăuzită de următoarele principii, va oferi oamenilor posibilităţi uimitoare în deceniile şi secolele care vor veni.

TEME DE CERCETARE

§ 1 Obiectiv de cercetare: *Obiectivul cercetării IA trebuie să fie acela de a crea o inteligenţa benefică, nu una nedirecţionată.*

§2 Finanţarea cercetării: *Investiţiile în IA trebuie însoţite de finanţarea cercetării care să asigure utilizarea benefică a acesteia, incluzând probleme delicate de informatică, ştiinţe economice, ştiinţe juridice, etică ţi sociologie, cum ar fi:*

(a) *Cum putem face ca sistemele IA din viitor să fie extrem de robuste, aţa încât să facă ceea ce vrem, fără erori de funcţionare ţi fără a fi vulnerabile în faţa hackerilor?*

(b) *Cum ne putem spori prosperitatea prin automatizare,fără a afecta resursele ţi voinţa oamenilor?*

(c) *Cum putem revizui sistemul nostru juridic pentru a fi mai corect ţi mai eficient, pentru a păstra ritmul cu IA ţi a gestiona riscurile*

asociate cu IA?

(d) *La ce set de valori trebuie aliniată IA ţi ce statut juridic ţi etic ar trebui să aibă?*

§3 Legătura dintre ştiinţă şi politică: *Trebuie să existe un dialog constructiv ţi sănătos între cercetătorii IA ţi oamenii politici.*

§4 Cultura cercetării: *Trebuiefăurită o cultură a cooperării, încrederii ţi transparenţei în rândul cercetătorilor ţi dezvoltatorilor IA.*

§5 Evitarea concurenţei neloiale: *Echipele care dezvoltă IA trebuie să coopereze activ pentru a împiedica recurgerea la soluţii facile care încalcă normele de siguranţă.*

ETICĂ ŞI VALORI

§6 Siguranţă: *Sistemele IA trebuie să fie sigure şi protejate în decursul întregului timp de viaţă al funcţionării lor, să poată fi verificate acolo unde sunt aplicabile şi unde sunt folosite.*

§7 Transparenţă faţă de eşec: *Dacă un sistem IA provoacă daune, trebuie să se poată stabili de ce.*

§8 Transparenţă juridică: *Orice implicare a unui sistem autonom în hotărâri judecătoreşti trebuie să ofere o explicaţie satisfăcătoare verificabilă de către o autoritate umană competentă.*

§9 Responsabilitate: *Proiectanţii şi constructorii sistemelor IA avansate sunt parte interesată în ce priveşte implicaţiile morale ale folosirii, folosirii greşite şi acţiunilor acestora, având responsabilitatea şi posibilitatea de a modifica acele implicaţii.*

§ io Alinierea valorilor: *Sistemele IA cu un grad înalt de autonomie trebuie proiectate aşa încât să poată fi asigurată alinierea obiectivelor şi comportamentelor lor cu valorile umane pe tot parcur- sulfuncţionării acestora.*

§ 11 Valorile umane: *Sistemele IA trebuie proiectate şi*

acţionate aşa încât să fie compatibile cu idealurile umane legate de demnitate, drepturi, libertăţi şi diversitate culturală.

§ 12 Confidenţialitate personală: *Oamenii trebuie să aibă dreptul de a accesa, gestiona şi controla datele pe care le generează, dată fiind capacitatea sistemelor IA de a analiza şi utiliza acele date.*

§ 13 Libertate şi confidenţialitate: *Aplicarea IA la datele personale nu trebuie să ştirbească libertatea reală sau percepută a oamenilor.*

§14 Beneficiu colectiv: *De tehnologiile IA trebuie să beneficieze cât mai rnulţi oameni cu putinţă.*

§15 Prosperitate partajată: *Prosperitatea economică creată de IA trebuie larg partajată, spre beneficiul întregii omeniri*

§16 Control uman: *Oamenii trebuie să aleagă cum şi dacă îşi deleagă deciziile sistemelor IA, pentru îndeplinirea propriilor lor obiective.*

§17 Anti-subminare: *Puterea conferită prin controlul unor sisteme IA avansate trebuie să respecte ţi să amelioreze, nu să submineze, procesele sociale şi civice de care depinde sănătatea societăţii.*

§18 Cursa înarmărilor IA: *Trebuie evitată o cursă a înarmărilor cu arme autonome letale.*

PROBLEME PE TERMEN LUNG

§19 Avertisment privind capacităţile: *Neexistând un consens în această privinţă, trebuie evitate presupunerile tranşante legate de limitele superioare ale capacităţilor viitoarelor IA.*

§ 20 Importanţă: *LA avansată poate reprezenta o projunda schimbare în istoria vieţii pe Pământ, şi trebuie planificată şi gestionată cu atenţia şi resursele corespunzătoare.*

§2t Riscuri: *Riscurile implicate de sistemele IA, în special cele care pot declanşa catastrofe sau ameninţă*

viaţa, trebuie supuse unor eforturi de prevedere şi diminuare pe măsura impactului lor potenţial § 22 Autoperfecţionare recursivă: *Sistemele IA proiectate pentru a se autoperfecţiona sau autoreplica recursiv, cu riscul de a deveni rapidfoarte numeroase sau foarte avansate, trebuie supuse unor măsuri stricte de siguranţă şi control.*

§23 Binele comun: *Superinteligenţa trebuie dezvoltată pentru a se afla numai în slujba unor idealuri etice larg împărtăşite şi în beneficiul întregii omeniri, nu doar al unui stat sau al unei organizaţii*

Lista de semnături a crescut spectaculos după ce am postat principiile online, iar acum conţine peste o mie de cercetători IA şi alte personalităţi de frunte. Dacă vrei şi tu să semnezi, o poţi face aici: http://futureoflife.org/ai-principles.

Am fost impresionaţi nu doar de nivelul ridicat al consensului asupra principiilor, ci şi de forţa acestora. Desigur, la prima vedere unele par de genul „Pacea, dragostea şi maternitatea sunt lucruri de preţ“, dar multe dintre ele îşi dovedesc eficacitatea atunci când formulezi negaţii ale lor. De pildă, „Superinteligenţa e imposibilă“ încalcă § 19, iar „Este pierdere de vreme cercetarea pentru a reduce riscul la care IA expune existenţa“ încalcă §2r.

După cum puteţi constata privind pe YouTube discuţia despre perspectivele pe termen lung,[11] Elon Musk, Stuart Russell, Ray

Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman şi Jaan Tallinnn au fost cu toţii de acord că superinteligenţa va fi probabil dezvoltată, iar cercetările legate de siguranţă sunt importante.

Sper ca principiile IA de la Asilomar să servească drept punct de pornire pentru discuţii mai detaliate, care să conducă în cele din urmă la strategii şi politici IA raţionale. în acest spirit, Ariei Conn, responsabil cu relaţiile publice la FLI, împreună cu Tucker Davey şi alţi membri ai echipei, au luat interviuri celor mai importanţi cercetători IA în legătură cu principiile şi felul în care le interpretează aceştia, în vreme ce David Stanley şi echipa lui internaţională de voluntari au tradus principiile în limbi de mare circulaţie.

OPTIMISM PRECAUT

După cum am mărturisit la începutul epilogului, sunt atât de optimist în privinţa viitorului vieţii cum n-am mai fost de mult. V-am împărtăşit povestea mea pentru a vă explica de ce.

Experienţele mele din ultimii câţiva ani mi-au sporit optimismul din două motive. în primul rând, am fost martor la o remarcabilă reunire a comunităţii IA pentru a accepta în mod constructiv provocările ce ne stau în faţă, adesea în colaborare cu gânditori din alte domenii. După întâlnirea de la Asilomar, Elon mi-a spus că îl uimea faptul că, în doar câţiva ani, dintr-o temă periferică, siguranţa IA a devenit una centrală, iar eu însumi sunt la fel de uimit. Nu doar problemele care se pun pe termen scurt, prezentate în capitolul 3, sunt acum subiecte de discuţie respectabile, dar şi superinteligenţa sau riscul la care e expusă existenţa, ca în principiile IA de la Asilomar. Acele principii nu puteau fi adoptate în Puerto Rico, cu doi ani mai devreme – cuvântul din scrisoarea deschisă care suna cel

mai înspăimântător a fost „capcane“.

îmi place să-i privesc pe oameni, iar în dimineaţa ultimei zile a conferinţei de la Asilomar am stat în marginea sălii şi m-am uitat

Figura 9.4 O comunitate tot mai numeroasă caută răspunsuri la Asilomar.

la participanţii care urmăreau o discuţie despre IA şi legi. M-a cuprins o senzaţie de căldură difuză şi am fost mişcat. Era cu totul altfel decât în Puerto Rico! Pe atunci, ţin minte că-i priveam pe cei mai mulţi din comunitatea IA cu un amestec de respect şi teamă – nu ca pe o echipă adversă, ci ca pe un grup pe care eu şi colegii mei preocupaţi de IA simţeam că trebuia să-l convingem. Acum era însă evident că făceam cu toţii parte din *aceeaşi* echipă. După cum aţi înţeles probabil citind această carte, tot nu am un răspuns la întrebarea cum ne putem făuri un viitor măreţ cu IA, aşa încât găsesc că e minunat să faci parte dintr-o comunitate în continuă creştere care caută răspunsul.

Al doilea motiv pentru care sunt mai optimist e acela că experienţa FLI a fost mobilizatoare. Ceea ce făcuse să-mi dea lacrimile la Londra a fost o senzaţie de inevitabil: ne-ar putea aştepta un viitor îngrijorător şi n-avem cum să ne

împotrivim. Următorii trei ani mi au risipit însă temerea fatalistă. Dacă o mână de voluntari neplătiţi pot influenţa dezbaterea pe care o consider cea mai importantă a epocii noastre, imaginaţi-vă ce am putea face cu toţii, dacă am lucra împreună!

Erik Brynjolfsson vorbea în cuvântarea sa de la Asilomar despre două tipuri de optimism. Există un optimism necondiţionat, ca în speranţa că soarele va răsări şi mâine, dar există şi ceea ce el numea „optimism precaut" – speranţa că se vor petrece lucruri bune dacă le plănuieşti cu atenţie şi te străduieşti pentru împlinirea lor. Acesta e optimismul pe care îl simt acum în privinţa viitorului vieţii.

Aşadar, ce poţi face *tu* pentru a influenţa în mod pozitiv viitorul vieţii, acum când intrăm în epoca inteligenţei artificiale? Din motive pe care le voi explica imediat, cred că primul pas este acela de a deveni un optimist precaut, dacă nu cumva eşti deja. Pentru a fi un bun optimist precaut, esenţial este să ajungi la o perspectivă pozitivă asupra viitorului. Când studenţii de la MIT vin să-mi ceară sfatul în privinţa carierei de urmat, încep de regulă prin a-i întreba unde se văd pe ei înşişi peste un deceniu. Dacă un student îmi răspunde „Am să fiu pesemne la spital, într-un salon de canceroşi, sau în cimitir, după ce m-a călcat un autobuz", nu-i acord mari şanse. Să-ţi priveşti viitorul doar în culori întunecate e o abordare catastrofală pentru planificarea carierei! Să-ţi dedici 100% din energie ca să eviţi bolile şi accidentele este calea către ipohondrie şi paranoia, nu către fericire. Aş vrea să-l aud pe student vorbind cu entuziasm despre obiectivele lui, după care putem discuta strategiile de a ajunge acolo, evitând capcanele.

Erik a subliniat că, în conformitate cu teoria jocurilor, perspectivele pozitive stau la baza unei mari părţi a colaborărilor în lume, de la căsătorii şi fuziuni de corporaţii până la hotărârea statelor independente de a

forma Statele Unite ale Americii. în fond, de ce să sacrifici ceva ce ai, dacă nu-ţi poţi închipui un câştig mai mare pe care l-ai obţine? Prin urmare, ar trebui să ne imaginăm un viitor pozitiv nu numai pentru noi înşine, dar şi pentru societate şi omenire. Cu alte cuvinte, avem nevoie de mai multă speranţă existenţială. După cum îmi reaminteşte adesea Meia, de la Frankenstein la Terminator, viziunile despre viitor în literatură şi film sunt predominant distopice. Altfel spus, noi, ca societate, ne plănuim viitorul cam la fel de rău ca ipoteticul meu student de la MIT. De aceea avem nevoie de mai mulţi optimişti precauţi. Şi tot de aceea v-am îndemnat în cartea de faţă să vă gândiţi la viitorul pe care vi-l *doriţi,* nu la viitorul de care vă temeţi, aşa încât să găsim obiective comune pe care să le plănuim şi cărora să le dedicăm eforturile noastre.

Am văzut în această carte că IA va aduce probabil cu sine posibilităţi măreţe şi provocări extreme. O strategie care ne va ajuta pesemne în practic toate provocările legate de IA este aceea de a acţiona împreună pentru a ameliora societatea umană *înainte* ca IA să ia avânt. Ar fi bine să-i educăm pe tineri pentru a face ca tehnologia să fie robustă şi benefică înainte să i se cedeze acesteia o mare parte din putere. Ar fi bine să ne modernizăm legile înainte ca tehnologia să le facă să devină desuete. Ar fi bine să rezolvăm conflictele internaţionale înainte de escaladarea unei curse a înarmărilor cu armament autonom. Ar fi bine să făurim o economie care asigură tuturor prosperitate înainte ca IA să poată amplifica inegalităţile. Am trăi mai bine într-o societate în care rezultatele cercetărilor privind siguranţa IA sunt puse în aplicare, nu ignorate. Şi, privind spre viitor la provocările legate de IAG supraumană, ar fi bine să cădem de acord măcar asupra câtorva standarde etice de bază, înainte de a pune maşini puternice să înveţe aceste standarde. într-o lume polarizată şi haotică, oamenii care au puterea de a

folosi IA în scopuri malefice vor fi tot mai motivaţi şi mai capabili s-o facă, iar echipele aflate în competiţie pentru construirea IAG vor fi supuse unei mari presiuni ca, în loc să coopereze în problemele de siguranţă, să trişeze. în concluzie, dacă putem edifica o societate umană mai armonioasă caracterizată prin cooperare pentru atingerea unor obiective comune, sporesc şansele ca revoluţia IA să urmeze un curs favorabil.

Cu alte cuvinte, una dintre cele mai bune căi de a ameliora viitorul vieţii este aceea de a ameliora ziua de mâine. Ai posibilitatea s-o faci în multe moduri. Poţi, desigur, vota la ume, spunând politicienilor ce crezi despre educaţie, confidenţialitate, armele autonome letale, şomaj tehnologic şi altele. Dar votezi şi zi de zi prin ceea ce alegi să cumperi, prin ştirile pe care alegi să le urmăreşti, prin ce alegi să împărtăşeşti celorlalţi, prin exemplul pe care alegi să-l dai. Vrei să întrerupi conversaţiile celorlalţi controlân-

du-le telefoanele inteligente, sau vrei ca tehnologia să fie folosită într-un mod planificat şi deliberat? Vrei ca tu să posezi tehnologia, sau vrei ca tehnologia să te posede pe tine? Ce ai vrea să însemne a fi om în epoca IA? Te invit să discuţi despre toate acestea cu cei din jurul tău- conversaţia nu e doar importantă, dar şi fascinantă.

Pe măsură ce modelăm epoca inteligenţei artificiale, suntem răspunzători pentru viitorul vieţii. Deşi la Londra am plâns, simt acum că nimic nu e inevitabil în privinţa acestui viitor, şi ştiu că e mult mai uşor să-l influenţezi decât îmi închipuiam. Viitorul nostru nu e săpat în piatră, urmând să ni se întâmple negreşit – noi suntem cei care îl construim. Să construim împreună unul rodnic!

Note

CAPITOLUL 1

1. „The Al Revolution: Our Immortality or Extinction?“, *Wait But Why* (27 ianuarie 2015), la http://waitbauwhy.com/2015/01/artificial-intelligen- ce-revolution-2.htlm.
2. Această scrisoare deschisă, „Research Priorities for Robust and Beneficial Artificial Intelligence“, poate fi găsită la http://futureoflife.org/ai-open-letter/.
3. Exemplu de alarmism clasic privind roboţii în mediile de informare: Ellie Zolfagharifard, „Artificial Intelligence «Could Be the Worst Thing to Happen to Humanity»“, *Daily Mail,* 2 mai 2014, http://tinyurl.com/hawkingbots.

CAPITOLUL 2

1. Observaţii despre originea termenului IAG: http://wp.goertzel.org/who- coined-the-term-agi.
2. Hans Moravec, „When Will Computer Hardware Match the Human Brain”, *Journal of Evolution and Technology* (1998), vol. r.
3. în figura care prezintă evoluţia puterii de calcul, datele dinainte de 2021 sunt luate din cartea lui Ray Kurzweil *Hout to Create a Mind,* iar datele ulterioare sunt calculate din referinţele de la https://en.wikipedia.org/wiki/FLOPS.
4. David Deutsch prezintă calculul cuantic drept o dovadă a

universurilor paralele în cartea sa *The Fabric of Reality (Allen* Lane, Londra, 1997). Opinia mea, după care universurile paralele cuantice reprezintă al treilea din cele patru niveluri ale multiversului, se găseşte în precedenta mea carte, *Our Mathematical Universe: My Questforthe Ultimate Nature of Reality (Knopf,* New York, 2014).

5- Pentru succesul echipei Google în descrierea imaginilor cu ajutorul IA vezi https://anciv.org/pdf/14114555 .pdf.

CAPITOLUL 3

1. Priviţi „Google DeepMind's Deep Q-Ieaming Playing Atari Breakout" pe YouTube la adresa: https://tinyurl.com/atariai.
2. Vezi Volodymyr Mnih *et al.,* „Human-Level Control Through Deep Reinforcement Learning", *Nature 518* {26 februarie 2015), disponibil online la http;//ti nyurl.com/ ataripaper.
3. Document video cu robotul Big Dog în acţiune: https://www.youtube. com/watch?v=W iczBcnXiWw.
4. Pentru reacţii la mutarea senzaţională a lui AlphaGO, vezi „Move 37!! Lee Sedol vs AlphaGo Match *2",* la; https://www.youtube.com/watch?v=JN- rXgpSEEIE.
5. Demis Hassabis prezentând reacţiile la AlphaGo ale unor jucători de go umani: https://www.youtube.com/watch7vsotJKzpNWZT4.
6. Pentru progresele recente în traducerea cu ajutorul maşinilor, vezi Gideon Lewis-Kraus, „The Great Al Awakening", *New York Times Magazine (ti* decembrie 2016), disponibil online la http://www.nytimes.c0m/2016/12/14/ magazine/the-great-ai-awakening.htmL GoogleTranslate e disponibil la https://translate.google.com.
7. Competiţia Provocarea Schemei Winograd: http://tinyurl.com/winogra- dchallenge.
8. Document video cu explozia rachetei Ariane 5: https://www.youtube.com/ watch?v=qnHn8WiEm6E.
9. Raportul privind explozia rachetei Ariane 5: http://tinyurl.com/arianeflop.
10. Raportul NASA privind eşecul sondei Mars Climate Orbiter: http^/tinyurl, com/marsflop.
11. Cea mai amănunţită şi consistentă relatare a cauzelor eşecului misiunii Mariner 1: http://tinyurl.com/marinerflop.

12.O descriere amănunţită a eşecului misiunii sovietice Phobos 1 poate fi găsită în Wesley T. Huntress Jr. şi Mikhail Ya. Marov, *Soviet Robots in the Solar System* (Praxis Publishing, New York, 2011), p. 308.

13. Despre felul în care un software neverificat a făcut ca Knight Capital să piardă 440 de milioane de dolari în 45 de minute: http://tinyurl.com/kni- ghtflopi şi http://tinyurL.com/knightflop2.

14. Raportul guvernamental privind „crahul fulger": „Findings Regarding the Market Events of May 6,2010" (30 septembrie 2010), la http://tinyurl.com/ flashcrashreport.

ij. Exemple de folosire a imprimantelor 3-D pentru construcţii (https://www. youtube.com/watch?v=SObzNdyRTBs), pentru dispozitive micromeca- nice (http://tinyurl.com/tinyprinter) şi pentru multe alte lucruri intermediare (https://www.youtube.com/watch?v=xVU4FLrsPXs).

16. Harta lumii pe care sunt trecute atelierele-laborator: https://www.fablabs. io/map.

17. Articol despre uciderea lui Robert Williams de către un robot industrial: http://tinyurl.com/williamsaccident.

18. Articol despre uciderea lui Kenji Urada de către un robot industrial: http:// tinyurl.com/uradaaccident.

19. Articol despre uciderea unui lucrător de la Volkswagen de către un robot industrial: http://tinyurl.com/baunatalaccident

20. Raport al guvernului SUA privind accidentele de muncă: https://www. osha.gov/dep/fatcat/dep_fatcat.html.

21. Statistica accidentelor rutiere: http://tinyurl.com/r0adsafety2 şi http:// tinyurl.com/roadsafety 3.

22. Despre primul accident al unei maşini Tesla autopilotate, vezi Andrew Buncombe, .Tesla Crash: Driver Who Died While on Autopilot Mode «Was Watching Harry Potter»", *independent (1* iulie 2016), http://tinyurl.com/ teslacrashstory. Raportul Administraţiei Naţionale pentru Traficul Rutier: http://tinyurl.com/teslacrashreport.

23. Amănunte despre catastrofa feribotului Herald of Free Enterprise găsiţi în R.B. Whittingham, *The Blame Machine: Why Human Error Cătun Acci-* dente(Elsevier, Oxford, 2004).

24. Documentar despre prăbuşirea zborului Air France 447:

https://www.you- tube.com/watch?v=dpPkp8OGQFI; raport despre accident http://tinyurl. com/af447ieport; analiză externă: http://tinyurl.com/thomsonarticle.

25. Raport oficial despre pana de curent din Statele Unite şi Canada: http:// tinyurl .corn/uscanadablackout.
26. Raport final al comisiei prezidenţiale despre accidentul de la Three Mile Island: http://www.threemileisland.org/downloads/i88.pdf.
27. Studiu olandez care arată călA poate rivaliza cu radiologii umani în diagnosticarea cancerului de prostată pe baza IRM: http://tinyurl.com/pTostate-ai.
28. Studiu Stanfbrd care arată că IA îi poate depăşi pe anatomopatologii umani în diagnosticarea cancerului la plămân: http://tinyurl.com/lungcancer-ai.
29. Investigaţie privind accidentele provocate de Therac-25: http://tinyurl. com/theracfailure.
30. Raport privind supradozele letale de la Institutul Oncologic Naţional din Panama: http://tinyurl .com/cobalt 6oacciden t.
31. Studiu privind accidentele provocate de chirurgia robotizată: https:// arxiv.0rg/abs/r507.035 r8.
32. Articol despre numărul deceselor provocate de malpraxisul medical în spitale: http://tinyurl.com/niedaccidents.
33. Yahoo a luat măsuri de siguranţă după ce în iox 3 a anunţat că hackerii au pătruns in peste un miliard (!) de conturi: https://www.nytimes.com/ 20i7/io/o3/technology/yahoo-hack-3-biIlion-users.html.
34. Articol din *New York Times* despre achitarea, iar apoi condamnarea asasinului care făcea parte din Ku Klux Klan: http://tinyurl.com/kkkacquittal.
35. Studiul efectuat de Danziger et al. în 2011 (http://www.pnas.org/con- tent/108/17/6889.(011), susţinând că judecătorii cărora le e foame sunt mai aspri, a fost criticat de Keren Weinshall-Margela şi John Shapard (http:// www.pnas.org/content/1o8/42/E833.full), dar Danziger *et aL* insistă că afirmaţiile lor rămân valabile (http://www.pnas.0rg/c0ntent/1 o8/42/E834.full).
36. *Pro Publica* afirmă că software-ul folosit pentru a prezice recidivismul nu e imparţial: http://tinyurl.com/robojudge.
37. Folosirea IRMf şi a altor tehnici de scanare cerebrală ca probă in procese este extrem de controversată, la fel şi încrederea în asemenea tehnici, deşi multe echipe susţin că

acurateţea e de peste 90%: http://joumal.frontier-sin.org/article/10.3 3 89/fpsyg. 2015.00709/full.

38. Reţeaua de televiziune PBS a făcut filmul *The Man Who Saved the Worlddes-* pre incidentul în care Vasili Arhipov a împiedicat de unul singur o lovitură nucleară sovietică: https?/www.youtube.com/watch?v=4VPY2SgyG5w.
39. Povestea lui Stanislav Petrov, care a considerat falsă avertizarea că urmează un atac nuclear american, a fost ecranizată în filmul *TheMan Who Saved the Worid* (a nu se confunda cu filmul cu acelaşi titlu din nota precedentă), iar ONU i-a acordat lui Petrov distincţia „Cetăţeanul Lumii“: https://www.youtube.com/watch7v-IncSjwWQHMo.
40. Scrisoare deschisă a cercetătorilor din IA şi robotică privind armele autonome: http://futureoflife.org/open-letter-autonomous-weapons/.
41. Un oficial american care pare să-şi dorească o cursă a înarmărilor LA http// tinyurLcom/ workquote.
42. Studiu despre inegalitatea averilor în Statele Unite începând din 1913: http// gabriel-zucman.eu/files/SaezZucman2oi5.pdf.
43. Organizaţia Oxfarn despre inegalitatea averilor http-J/tinyurLcom/oxfamioi 7.
44. Pentru o prezentare a ipotezei că inegalitatea e indusă de tehnologie, vezi Erik Brynjolfsson şi Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time gfBrilliant Technologies* (Norton, New York, 2014).
45. Articol din *The Atlantic* despre scăderea salariilor celor mai puţin instruiţi: http://tinyurl.com/wagedrop.
46. Datele din grafic sunt luate din Facundo Alvaredo, Anthony B. Atkinson, Thornas Piketty, Emmanuel Saez şi Gabriel Zucman, *The Worid Wealthand Income* Dafa/>a$e(http://www.wid.world, 31/10/2016), incluzând câştigurile de capital
47. Prezentare a lui James Manyika: http://futureoflife.org/data/PDF/james_ manyika.pdf.
48. Previziuni privind viitorul automatizării joburilor elaborate de Universitatea Oxford (http://tinyurl.com/automationoxford) şi de McKinsey (http://tinyurl.com/automationmckinsey).
49. Document video cu un bucătar robot

https://www.youtube.com/watch?- v-fE6i2OO6Y6s.

50. Marin Solacic a examinat aceste opţiuni la seminarul din 2016 „Computere Gone Wild: Impact and Implications of Developments in Artificial Intelligence on Society“: http://futureoflife.org/2016/05/06/computere-gone-wild/.
51. Propunerile lui Andrew McAfee pentru crearea mai multor joburi: http:// futureoflife.org/data/PDF/andrew_mcafee.pdf.
52. în plus faţă de numeroasele articole academice susţinând că şomajul tehnologic „de data asta e altfel“, documentul video „Humans need not apply“ prezintă pe scurt aceeaşi idee: https://www.youtube.com/watch7vs $_{7}$Pq-S$_{5}$57XQU.
53. Oficiul Statelor Unite pentru Statistica Muncii: http://www.bls.gov/cps/ cpsaatn.htm.
54. Argument că „de data asta e altfel“ pentru şomajul tehnologic: Federico Pistono, *Robots Will Steal Your Job, but That's* OK (2012), http://robotswills- tealyourjob.com.
55. Evoluţia şeptelului cabalin american: http://tinyurl.com/horsedecline.
56. Meta-analiză arătând cum afectează şomajul starea de bine Maike Luhmann *et al,* „Subjective Well-Being and Adaptation to Life Events; A Meta-Analysis“, *Journal of personality and social psychology 102,* nr. 3 (2012), p. 592, disponibil online la https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3 2897 59.
57. Studii privind sursele senzaţiei de bunăstare Angela Duckworth, Tracy Steen şi Martin Seligman, „Positive Psychology in Clinical Practice“, Annua/ *Review of Clinical Psychology* 1 (2005), pp. 629-651, online la http://tinyurl. com/wellbeingduckworth; Weiting Ng şi Ed Diener, „What Matters to the Rich and the Poor? Subjective Well-Being, Financial Satisfaction, and Postmaterialist Needs Across the World”, *Journal of personality and social psychology* 107, nr. 2 (2014), p. 326, online la http://psycnet.apa.org/journals/psp/i07/a/326; Kirsten Weir, „More than Job Satisfaction“, *Monitor on Psychology* 44, nr. 11 (decembrie 2013), online la http://www.apa.org/ monitor/2013/12/job-satisfaction.aspx.
58. înmulţind circa io^{11} neuroni cu circa io^{4} conexiuni pe neuron şi cu circa 0 (ro°) activare pe neuron în fiecare secundă, ar rezulta că aproximativ io^{1}5 FLOPS (i

petaFLOPS) sunt de ajuns pentru a simula un creier uman, dar rămân nelămurite încă multe lucruri, între care ritmul activărilor şi întrebarea dacă nu cumva părţile mici ale neuronilor şi sinapselor trebuie şi ele simulate. Informaticianul Dharmendra Modha de la IBM a estimat că e nevoie de 38 de petaFLOPS (http://tmyurl.c01n/javln43), în vreme ce specialistul în neuroştiinţe Henry Markram a estimat că e nevoie de circa 1000 de petaFLOPS (http://tinyurl.coTn/6rpohqv). Cercetătorii IA Katja Grace şi Paul Christiano au susţinut că aspectul cel mai costisitor al simulării nu e calculul, ci *comunicarea,* iar şi această sarcină poate fi în linii mari îndeplinită de cele mai bune supercalculatoare actuale: http-J/aiimpacts.org/about.

59. Pentru o estimare interesantă a puterii de calcul a creierului uman: Hans Moravec, „When Will Computer Hardware Match the Human Brain“, *Joumal ofEvolution and Technology,* voi. 1 (1998).

CAPITOLUL 4

1. Pentru un document video cu prima pasăre mecanică, vezi Markus Fischer, „A Robot That Flies Like a Bird“, Conferinţă TED, iulie 2011, la https://www.ted.com/talks/a_robot_that_flies_like_a_bird.

CAPITOLUL 5

r. Ray Kurzweil, *The Singulari® Is Near* (Viking Press, New York, 2005).
2. Scenariul lui Ben Goertzel „Nanny AI“ e prezentat la https://wiki. lesswrong.com/wiki/N anny_AI.
3. Pentru o discuţie despre relaţia dintre maşini şi oameni, vezi Benjamin Wallace-Wells, „Boyhood“, *New York Magazine* (20 mai 2015), online la http://tinyurl.com/ aislaves.
4. Crima împotriva minţii e discutată în cartea lui Nick Bostrom *Superintelligence,* şi mai detaliat în acest articol recent: Nick Bostrom, Allan Dafoe şi Carrick Flynn, „Policy Desiderata in the Development of Machine Superintelligence“ (2016), http://www.nickbostrom.com/papers/aipolicy.pdf.
5. Matthew Schofield, „Memories of Stasi Color German’s View of U.S. Surveillance Programs“, *McClatchj) DC Sureau* (26

iunie 2013), online la http://www.mcclatchydc.com/news/nation-world/national/article24750439.ht1nL

6. Pentru reflecţii pătrunzătoare privind felul în care oamenii pot fi îndemnaţi să acţioneze pentru a se ajunge la rezultate pe care nimeni nu şi le doreşte, recomand „Meditations on Molloch“, http://slatestarcodex. com/2014/07 / 30/meditations-on-moloch.
7. Pentru o listă a cazurilor in care războiul nuclear a fost cât pe ce să fie declanşat accidental, vezi Institutul pentru Viitorul Vieţii, „Accidental Nuclear War A Timeline of Close Calls”, online la http://tinyurLcoin/nukeoops.

8. Pentru despăgubiri acordate victimelor testelor nucleare americane, vezi site-ul Departamentului de Justiţie al Statelor Unite, „Awards to Date 4/24/2015“, la https://www.justice.gov/civil/awards-date-04242015.
9. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) A tack,* aprilie 2008, disponibil online la http://www. empcommission.org/docs/A2473-EMP_Commission-7MB.pdf.
10. Cercetări independente ale oamenilor de ştiinţă americani şi sovietici i-au prevenit pe Reagan şi Gorbaciov în privinţa riscului iemii nucleare: P.J. Crutzen şi J.W. Birks, „The Atmosphere After a Nuclear War: Twilight at Noon”, *Ambio* 11, nr. 2/3 (1982), pp. 114-125; R.P. Turco, O.B. Toon, T.P. Ackerman, J.B. Pollack şi C. Sagan, „Nuclear Winter: Global Consequences of Multiple Nuclear Explosions”, *Science 222* (1983), pp. 1283-1292; V.V. Aleksandrov şi G.L. Stenchikov, „On the Modeling of the Climatic Consequences of the Nuclear War”, *Proceeding on Applied Mathematics* (Moscova, Centrul de Calcul al Academiei de Ştiinţe a URSS) 21; A. Robock, „Snow and Ice Feedbacks Prolong Effects of Nuclear Winter“, *Nature* 310 (1984), pp. 667-670.

ir. Calcul al efectelor războiului nuclear global asupra climei: A. Robock, L Oman şi L. Stenchikov, „Nuclear Winter Revisited with a Modem Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences“, *Journal of Geophysical Research* 12 (2007), P.D13107.

CAPITOLUL 6

1. Pentru informaţii suplimentare, vezi Anders Sandberg, „Dyson Sphere FAQ“, la http://tinyurl.com/dysonsph.
2. Articolul lui Freeman Dyson despre sferele care-i poartă numele: Freeman Dyson, „Search for Artificial Stellar Sources of Infrared Radiation“, *Science,* vol. 131 (1959), pp. 1667-1668.
3. Louis Crane şi Shawn Westmoreland explică motorul cu găuri negre propus de ei în „Are Blach Hole Starship Possible?“, la http://arxiv.org/ pdf/o9o8.i8o3-pdf.
4. Pentru un grafic produs de CERN care trece în revistă particulele elementare cunoscute, vezi http://tinyurl.com/cemparticles.
5. Acest document video cu un prototip Orion nenuclear ilustrează ideea de propulsie cu bombe nucleare: https://www.youtube.com/watch?v=E- 3LxxzVAYi8.
6. 0 prezentare accesibilă a navigaţiei pe bază de laser: Robert L. Forward, „Roundtrip Interstellar Travel Using Laser-Pushed Lightsails“, *Journal of Spacecrafts and Rockets* 21, nr. 2 (martie-aprilie 2014), disponibil online la http://www.lunarsail.com/LightSail/rit-i.pdf.

7- Jay Olson analizează expansiunea cosmică a civilizaţiilor în „Homogenous Cosmology with Aggressively Expading Civilizations“, *Classical and Quantum Gravity* 32 (2015), disponibil online la http://arxiv.org/abs/14iM359

8. Prima analiză ştiinţifică a viitorului nostru îndepărtat Freeman J. Dyson, „Time Without End: Physics and Biology in an Open Universe”, *Reviews of Modem Physics* 51, nr. 3 (1979), p. 447, disponibil online la http://blog. regehr.org/ extra_files/dy son.pdf.
9. Formula lui Seth Lloyd menţionată mai sus ne spune că efectuarea unei operaţii de calcul în intervalul de timp ţ consumă o energie *Ei h/4-c,* unde heste constanta lui Planck. Dacă vrem să efectuăm *N* operaţii una după alta (în serie) în timpul T, atunci r = *T7N,* deci *E/N> hN/tT,* de unde rezultă că putem efectua *N<t ETih*operaţii în serie folosind energia Eîn timpul *T.* Dacă împărţim energia între n calcule paralele, ele pot fi efectuate mai lent şi mai eficient, aşa încât *N<* 2 *fEfn/ h.* Nick Bostrom estimează că simularea unei vieţi umane de 100 de ani necesită aproximativ N-10^{27} operaţii.
10. Un raţionament care arată că originea vieţii presupune un accident extrem de rar, plasând vecinii noştri cei mai

apropiaţi la o distanţă de peste io[1000] metri, poate fi găsit în acest document video al fizicianului şi astrobi- ologului de la Princeton Edwin Tumen „Improbable Life: An Unappealing but Plausible Scenario for Life's Origin on Earth", la https://www.youtube. com/watch?v=Bt6n6Tu 1 beg.

11. Eseu al lui Martin Rees despre căutarea inteligenţei extraterestre: https:// www.edge.org/annual-question/zo16/response/26665.

CAPITOLUL 7

1. O prezentare accesibilă a cercetărilor lui Jeremy England privind „adaptarea determinată de disipare" poate fi găsită în Natalie Wolchover, *Js* New Physics Theory of Life", *Scientific American* (28 ianuarie 2014), disponibil online la https://www.scientificamerican.com/article/a-new-phy- sics-theory-of-life/. Cartea lui Ilya Prigogine şi Isabelle Stengers *Order Out of Chaos: Man's New Dialogue with Nature* (Bantam, New York, 1984) stabileşte în mare măsură fundamentele acestor cercetări.
2. Pentru bazele fiziologice ale sentimentelor, vezi: William James, *Principles of Psychology* (N ew York, Henry Holt & Co., 1890); Robert Omstein, *Evolution of Consciousness: The Origins of the* Wap *We Think* (New York, Simon & Schuster, 1992): Antonio Damásio, *Descartes'Error Emotion, Reason, and the Human Brain* (Penguin, New York, 2005) (ed. rom. *Eroarea lui Descartes. Emoţiile, raţiune şi creierul uman,* Humanitás, Bucureşti, 2005^ Antonio Damásio, *Self Comes to Mind: Constructing the Conscious Bram(VintAge.* New York, 2012) (ed. rom. *Şinele. Construirea creierului inteligent,* Humanitás, Bucureşti, 2019].

3. Eliezer Yudkowsky a analizat alinierea obiectivelor unei IA prietenoase nu cu obiectivele noastre actuale, ci cu *voinţa noastră extrapolată coerent (coherent extrapolated volition-CEV).* în linii mari, aceasta e definită drept ceea ce ar vrea o versiune idealizată a noastră dacă ar şti mai multe şi ar gândi mai repede, reprezentând mai curând ceea ce am vrea noi să fim. Yudkowsky a început să critice CEV la scurt timp după ce a publicat articolul despre ea (http://intelligence.org/fiIes/CEV.pdf), pe motiv că e greu de operat cu ea şi nu e limpede dacă poate converge către ceva bine definit.

4. în abordarea învăţării prin consolidare inversă, o idee

esenţială este că IA nu încearcă să maximizeze satisfacţia îndeplinirii obiectivelor pentru sine, ci pentru proprietarul ei uman. Ea este deci îndemnată să fie precaută atunci când nu ştie clar ce vrea proprietarul, şi să se străduiască să afle. Ar trebui de asemenea să nu aibă de obiectat dacă proprietarul o scoate din funcţiune, fiindcă asta înseamnă că i-a înţeles greşit dorinţa.

5. Articolul lui Steve Omohundro despre apariţia obiectivelor unei LA poate fi găsit la http://tinyurl.com/omohundro2008. El a fost publicat iniţial în *Artificial GeneralIntelligence 2008: Proceedings aftheFirstAGIConference,* ed. Pei Wang, Ben Goertzel şi Stan Franklin (IOS, Amsterdam, 2008), pp. 483-492,

6.0 carte despre ce se întâmplă când inteligenţa e folosită pentru a se supune orbeşte ordinelor, iară să-şi pună probleme privind baza lor etică: Hanna Arendt, *Eichmann in Jerusalenv* A *Report on the Banality of* £w?(Penguin, New York, 1963). O dilemă înrudită se leagă de propunerea lui Eric Drexler (http://www.fhi.ox.ac.uk/reports/2015-3.pdf) de a ţine superinteligenţa sub control prin compartimentarea ei în părţi simple, aşa încât nici una dintre acestea să nu înţeleagă imaginea de ansamblu. Dacă metoda ar funcţiona, s-ar obţine un instrument incredibil de puternic fără vreo busolă morală intrinsecă, care ar pune în aplicare toate capriciile proprietarului ei, fără scrupule de ordin moral. Ar semăna cu o birocraţie compartimentată din- tr-o dictatură distopică: o parte construieşte arme fără să ştie că ele vor fi folosite, alta execută deţinuţi fără să ştie că au fost condamnaţi etc.

7.0 variantă modernă a Regulii de Aur este ideea lui John Rawls că o situaţie ipotetică e acceptabilă dacă nimeni n-ar schimba-o fără să ştie dinainte cine se va confrunta cu ea.

8. De pildă, coeficientul de inteligenţă al multor fucţionari de rang ¡nalt ai lui Hitler pare să fi fost destul de ridicat. Vezi „How Accurate Were the 1Q Scores of the High-Rankîng Third Reich Officials Tried at Nuremberg?“, *Quora,* disponibil la http://tinyurl.com/nurembergiq.

CAPITOLUL 8

1. Articolul despre conştiinţă al lui Stuart Sutherland e foarte amuzant: *Macmillan Dicţionar? ofPsychology (Macmiliün,*

bondar, 1989).

2. Erwin Schrodinger, unul dintre părinţii mecanicii cuantice, a făcut această remarcă în cartea sa Mind *and Matter* gândindu-se la *trecut-* ţi la ce s ar fi întâmplat dacă viaţa conştientă n-ar fi evoluat Pe de altă parte, ascensiunea IA face să apară în mod logic posibilitatea de a avea o piesă care să se joace în faţa unei săli pustii în *viitor.*
3. *Stanford Encyclopedia of Philosophy* dă diferite definiţii ţi întrebuinţări ale cuvântului .conştiinţă“: http://tinyurl.com/stanfordconsciousness.
4. Yuval Noah Harari, *HomoDeus:A Brief History* o/Tcnwmw (Harper-Collins, New York), p. 116 [ed, rom. *Homo Deus. Scurtă istorie a viitorului,* Polirom, Iaşi, 2018].
5. 0 excelentă prezentare a Sistemului 1 ţi a Sistemului 2: Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus & Giroux, New York, 2011).
6. Vezi ChristofKoch, *TheQuestforConsciousness: A Neurobiological Approach* (W.H. Freeman, New York, 2004).
7. Suntem pesemne conţtienţi doar de o infimă parte (10-50 biţi) din informaţia care intră în creierul nostru în fiecare secundă: K. Küpfmüller, „Nachrichtenverarbeitungim Menschen“,în *TaschenbuchderNachrichtenve- rarbeitung,* ed. K.Steinbuch, (Springer-Verlag, Berlin, 1962), pp. 1481-1502; T. Norretranders, *The User Illusion: Cutting Consciousness Down to Size* (Viking, New York, 1991).
8. MichioKaku, *The Future ofthe Mind: The Scientific Quest to Understand, Enhance, andEmpower the Mind* (Doubleday, New York, 2014k Jeff Hawkins şi Sandra Blakeslee, *On Intelligence (Times* Books, New York, 2007); Stanislas Dehaene, Michel Kerszberg ţi Jean-Pierre Changeux, „A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks”, *Proceedings of the National Academy of Sciences 95* (1998),pp. 14529-14534.
9 Document video despre celebrul experiment al lui Penfield „simt miros de pâine prăjită“: https;//www.youtube.com/watch?v=mSN86kphl68. Detalii despre cortexul somatosenzorial: Elaine Marieb ţi Katja Hoehn, *Anatomy & Physiology,* ed. a Ill-a (Pearson, Upper Saddle River, NJ, 2008), PP- 391-395-
ro Studiul corelativelor neuronale ale conştiinţei a intrat în ultimii ani în centrul atenţiei- vezi, de pildă, Geraint Rees,

Gabriel Kreiman şi Christof Koch, „Neural Correlates of Consciousness in Humans“, *Nature Reviews Neuroscience* 3 (2002), pp. 261-270, ţi Thomas Metzinger, *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, Cambridge, MA, 2000).

r i. Cum funcţionează suprimarea continuă a instantaneelor: Christof Koch, *The Quest for Consciousness:* A *Neurobiological ApproachfW.H.* Freeman, New York, 2004k Christof Koch şi Naotsugu Tsuchiya, „Continuous Flash Suppression Reduces Negative Afterimages“, *Nature Neuroscience 6* (2005), pp. 1096-1101,

12. Christof Koch, Marcello Massimini, Melanie Boly şi Giulio Tononi, „Neural Correlates of Consciousness: Progress and Problems”, *Nature Reviews Neuroscience* 17 (2016), p. 307.
13. Vezi Koch, *The Quest for Consciousness,* p. 260 şi discuţia din *Stanford Encyclopedia of Philosophy,* http://tinyurl.com/consciousnessdelay.
14. Despre sincronizarea percepţiei conştiente: David Eagleman, *The Brain:* The Story o/You(Pantheon,New York, 2015) [ed. rom. *Creierul. Povestea noastră,* Humanitás, Bucureşti, 2018] şi *Stanford Encyclopedia of Philosophy,* http:// tinyurl.com/consciousnesssync.
15. Benjamin Libet, Mind Time: *The Temporal Factor in Consciousness* (Harvard University Press, Cambridge, MA, 2004); Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze şi John-Dylan Haynes, „Unconscious Determinants of Free Decisions in the Human Brain”, *Nature Neuroscience* 11 (2008), pp. 543-545, online la http://www.nature.com/neuro/joumal/v11/n5/full/ nn.2112.html.
16. Exemple de abordări teoretice recente ale conştiinţei: Daniel Dennett, *Consciousness Explained (tick* Bay Books, 1992); Bemard Baars, *În the Theater of Consciousness: The Workspace ofthe Mind* (Oxford University Press, New York, 2001); Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (W.H. Freeman, New York, 2004); Gerald Edelman şi Giulio Tononi, A *Universe Of Consciousness How Matter Becomes Imagination: How Matter Becomes Imagination* (Hachette, New York, 2008); Antonio Damásio, *Self Comes to Mind: Constructing the Conscious* Brain (Vintage, New York, 2012); Stanislas

Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Viking, New York, 2014): Stanislas Dehaene, Michel Kerszberg şi Jean- Pierre Changeux, „A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks", *Proceedings of the National Academy of Sciences* 95 (1998), pp. 14529-14534; Stanislas Dehaene, Lucie Charles, Jean-Rémi King şi Sébastien Marti, „Toward a Computational Theory of Conscious Processing", *Current Opinion in Neurobiology* 25 (2014), pp. 760-784.

17. Pentru o discuţie aprofundată privind diferitele întrebuinţări ale termenului „emergenţă" în fizică şi filozofie: David Chalmers, http://cse35Zi. artifice.cc/Chalmers-Emergence.pdf.
18. Argumentul meu conform căruia conştiinţa este ceea ce simte informaţia atunci când e procesată în anumite moduri complexe: https://arxiv.org/abs/ physics/0510188; https://arxiv.0rg/abs/0704.0646; Max Tegmark, *Our Mathematical Universe (Knopf, New* York, 2014). David Chalmers exprimă o idee asemănătoare în cartea sa din 1996 *The Conscious Mind:* „Experienţa este informaţia privită din interior; fizica este informaţia privită din exterior."
ic Adenauer Casali et *al,* „A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior", *Science Translational Mediane* 5 (2013), 1981a! 05, online la http://tinyurl.com/zapzip.
20. Teoria informaţiei integrate nu se aplică sistemelor continue: https://arxiv. org/abs/i 401.1219; http://iournalfrontiersin.org/article/10.3 jög/fpsyg.zo 14. 00063/full; https://arxiv.0rg/abs/1601.02626.
21. Interviu cu Clive Wearing, a cărui memorie pe termen scurt durează doar circa 30 de secunde: https://www.youtube.com/watch7vsWmzU47i2xgw.
22. Critica lui Scott Aaronson la adresa IIT: http://www.scottaaronson.com/ blog/?p-i799.
23. Critica lui Cerrullo la adresa UT, susţinând că integrarea nu e o condiţie suficientă pentru conştiinţă: http://tinyurl.com/cerrullocritique.
24. Predicţia UT că oamenii simulaţi vor fi zombi: http://rstb.royalsocietypu- blishing.org/content/ 370/1668/20140167.
25. Critica lui Shanahan la adresa IIT: http://arxiv.org/ftp/arxiv/papers/1504/ 1504.05696.pdf.

26. Vedere oarbă: http://tinyurl.com/blindsight-paper.
27. Poate că suntem conştienţi doar de o parte minusculă (10-50 biţi) din informaţia care intră în creierul nostru în fiecare secundă: Kiipfmuller, „Nachrichtenverarbeitung im Menschen“; Norretranders, *The User Illusion.*
28. Argumente pro ţi contra „conştiinţei föra acces“: Victor Lamme, „How Neuroscience Will Change Our View on Consciousness”, *Cognitive Neuroscience* (2010), pp. 204-220, online la http7/www.tandfonline.com/doi/abs/10.1080/17588921003731586.
29. „Selective Attention Test“, https://www.youtube.com/watch7v-vJG698 UzMvo.
30. Vezi Lamme, „How Neuroscience Will Change Our View on Consciousness”.
31. Acesta ţi alte subiecte înrudite sunt discutate în cartea lui Daniel Dennett *Consciousness Explained.*
32. Vezi Kahneman, *Thinking, Fast and Slow.*
33. *Stanford Encyclopedia of Philosophy* despre liberul-arbitru: https://plato.stan- ford.edu/entries/freewill.
34. Document video cu Seth Lloyd explicând de ce o IA va simţi că are liber-ar- bitru: https://www.youtube.com/watchTvsEpj3DF8JDWk.
35. Vezi Steven Weinberg, *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature* (Pantheon, New York, 1992) [ed. rom. *Visul unei teorii finale. în căutarea legilor ultime ale naturii,* Humanitas, Bucureşti, 2010].
36. Prima analiză ştiinţifică a viitorului nostru îndepărtat Freeman J. Dyson, „Time Without End: Physics and Biology in an Open Universe“, *Reviews of Modern Physics* 51, nr. 3 (1979), p. 447, disponibil online la http://tinyurl.com/dysonfuture.

EPILOG

1. Scrisoarea deschisă (http://futureoflife.org/ai-open-letter/) apărută în urma conferinţei din Puerto Rico susţinea că studierea modului de a face sistemele IA robuste şi benefice este deopotrivă necesară şi urgentă, şi că există direcţii concrete de cercetare care pot fi urmate încă de pe acum, aşa cum e exemplificat în acest document despre priorităţile în cercetare; http://futureoflife.org/data/documents/research_priorities.pdf.
2. Interviul meu cu Elon Musk despre siguranţa IA: https://www.youtube. com/watch?v=rBwoeoZTY-g
3. Compilaţie video cu aproape toate încercările de aterizare ale rachetei SpaceX, culminând cu prima amerizare reuşită: https://www.youtube. com/watch?v=AllaFzIPaG4.
4. Elon Musk anunţă pe tweeter bursele pentru siguranţa IA https://twitter.
   com/elon musk/status/5 55743387056226304.
5. Elon Musk anunţă pe tweeter scrisoarea noastră deschisă în favoarea siguranţei IA https://twitter.com/eionmusk/status/554320532T33650452.
6. Erik Sofge îşi bate joc de ştirile panicarde declanşate de scrisoarea noastră
   deschisă: „An Open Letter to Anyone Tricked into Fearing Artificial Intelligence“ *(Popular Science,* 14 ianuarie 2015), online la http://www.po- psci.com/open-letter-everyone-tricked-fearing-ai.
7. Elon Musk anunţă pe tweeter donaţia către Institutul pentru Viitorul Vieţii şi către cercetătorii în siguranţa IA https://twitter.com/elonmusk/ status/555 743387056226304.
8. Mai multe despre parteneriatul pentru IA în beneficiul oamenilor şi so cietăţii găsiţi la https://www.partnershiponai.org.
9. Câteva exemple de rapoarte recente conţinând opinii despre IA: htt- tinyurl.com/stanfordai, http://tinyurl.com/obamaAIreport, ht tinyurl.com/AIjobsreport, http://standards.ieee.org/develop/indco’ eadvi .pdf, http://tinyurl.com/roboticsmap.

ro. între principiile care n-au fost adoptate, unul dintre preferate' fost următorul: ^Avertisment privind conştiinţa: Neexistând r în această privinţă, trebuie evitate presupunerile tranşante • punsul la întrebarea dacă o IA avansată ar avea sau ar trebui inţă sau sentimente.“ La capătul mai multor iteraţii, cuvânt „conştiinţă” a fost înlocuit cu „experienţe subiective” – piu a primit doar 88% din voturi, cu puţin sub pragul
11. Discuţie despre superinteligenţă cu Elon Musk şi altr http://tinyurl.com/asilomarAI.

Cuprins

Mulţumiri..... 7

Preludiu; Povestea Echipei Omega..... 11

1. Sunteţi invitaţi la cea mai importantă dezbatere a epocii noastre... 31

2. Materia devine inteligentă..... 60

3. Viitorul apropiat: progrese, erori, legi, arme şi joburi..... 95

4. Explozia inteligenţei?..... 152

5. Consecinţe: următorii 10 000 de ani..... 180

6. Zestrea noastră cosmică: următorul miliard de ani şi mai târziu.... 226

7. Obiective..... 278

8. Conştiinţa..... 314

Epilog: Povestea Echipei FU..... 353

SERIA ŞTIINŢĂ

este coordonată de Vlad Zografi

Alte cărţi de ştiinţă la Humanitás:

CARLO ROVELL1
Realitatea nu e ceea ce pare

CARLO ROVELLI
Ordinea timpului

[IM BAGGOTT
Origini:
Povestea ştiinţifică a creaţiei

PAOLO ZELL1N1
Matematica zeilor şi algoritmii oamenilor

STEVEN WEINBERG
Lumea explicată:
Descoperirea ştiinţei moderne

[ACOB BRONOWSKI
Ascensiunea omului

FRANS DE WAAL
Bonobo şi ateul:
în căutarea umanismului printre primate

Pe copertă:
© Guliver/iSiockphoto/Jirsak

Noi, cu toţii ~ nu doar oamenii de ştiinţă, industriaşii şi generalii ar trebui să ne întrebăm ce putem face pentru a profita de beneficiile viitoarei inteligenţe artificiale şi a evita riscurile. Aceasta e cea mai importantă dezbatere a epocii noastre, iar incitanta carte a lui Tegmark vă îndeamnă să participaţi la ea. —STEPHEN HAWKING

Fizicianul Max Tegmark, profesor la Institutul Tehnologic din Massachusetts, explorează posibilităţile oferite de inteligenţa artificială nu doar în viitorul imediat, dar şi în perspectiva îndepărtată a colonizării cosmosului. *Viaţa* 3.0 ne invită să medităm asupra celor mai delicate probleme de ordin social, militar, juridic şi etic generate de dezvoltarea inteligenţei artificiale, iar ideea centrală a lui Max Tegmark este că avem capacitatea şi datoria morală de a influenţa evoluţia vieţii.

ISBN 978-973-50-6402-0

* *

Aceasta a fost povestea Echipei Omega. Restul cărţii spune o altă poveste – una care n-a fost încă scrisă: povestea viitorului nostru cu IA. Cum aţi vrea să se desfăşoare? S-ar putea întâmpla ceva care să semene oarecum cu povestea Omega? Aţi vrea să fie aşa? Lăsând deoparte speculaţiile privind IA supraumană, cum aţi vrea să înceapă povestea noastră? Cum aţi vrea să influenţeze IA locurile de muncă, legile şi armamentul în deceniul următor? Privind mai departe, cum aţi scrie sfârşitul ei? Această poveste are într-adevăr proporţii cosmice, odată ce implică nici mai mult, nici mai puţin decât soarta vieţii în universul nostru. Şi e o poveste pe care noi o vom scrie.

*

* *

Vom vedea imediat care sunt punctele vulnerabile din planul omegilor, dar presupunând pentru o clipă că ar funcţiona, ce crezi